UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXX

Inference Following Two-Stage Randomization Designs with Survival Endpoints

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Giuliana Cortese

Dottorando: Sifiso Vilakati

31 October 2017

Abstract

Treatment of complex diseases such as cancer, HIV, leukemia and depression usually follows complex treatment sequences. In two-stage randomization designs, patients are randomized to first-stage treatments, and upon response, a second randomization to the second-stage treatments is done. The clinical goal in such trials is to achieve a response such as complete remission of leukemia, 50% shrinkage of solid tumor or increase in CD4 count in HIV patients. These responses are presumed to predict longer survival.

The focus in two-stage randomization designs with survival endpoints is on estimating survival distributions and comparing different treatment policies. In this thesis, we make contributions in these two areas. A simulation study is conducted to compare three non-parametric methods for estimating survival distributions. A parametric method is proposed for estimating survival distributions in time-varying SMART designs. The proposed estimator is studied using simulations and also applied to a clinical trial dataset. Thirdly, we propose a method for comparing different treatment policies. The new method works well even if the survival curves from the treatment policies cross. Simulation studies show that the new method has better statistical power than the weighted log-rank test in cases where survival curves cross. The last part of this thesis focuses on analyzing adverse events data from two-stage randomization designs. We develop a methodology for analyzing adverse events data in the competing risk setting which has been applied to a leukemia clinical trial dataset.

Sommario

Il trattamento di malattie complesse come cancro, AIDS, leucemia e depressione richiedono solitamente l'applicazione sequenziale di terapie complesse multiple. Nei disegni randomizzati a due stadi, inizialmente i pazienti sono randomizzati al primo stadio di trattamenti, e successivamente, sulla base della risposta al trattamento, i pazienti sono randomizzati ad un secondo stadio di trattamenti. In questi studi randomizzati, l'obiettivo clinico è quello di ottenere una risposta all'intero piano di trattamento, come per esempio la remissione completa dalla leucemia, la riduzione del 50% di un tumore solido, o l'aumento della proteina CD4 in pazienti con infezioneda HIV. Si presume che la risposta al trattamento possa predire una sopravvivenza più lunga.

Nei disegni randomizzati a due stadi che coinvolgono una risposta sul tempo di sopravvivenza, l'interesse principale è rivolto sia a stimare le distribuzioni di sopravvivenza sia a confrontare le variepolitiche di trattamento. La tesi di dottorato fornisce contributi di ricerca su questi due aspetti. È stato condotto uno studio di simulazione per confrontare diversi metodi non parametriciesistenti in letteratura per la stima delle distribuzioni di sopravvivenza. È stato proposto un metodo parametrico per stimarele distribuzioni di sopravvivenza in disegni randomizzati a due stadi di tipo SMART tempo-dipendente ("time-varying SMART designs"). Lo stimatore proposto è stato verificato tramite studi di simulazione ed è stato applicato a dati relativi a prove cliniche di trattamenti per la leucemia.

In terzo luogo, è stato proposto un metodo di verifica di ipotesi per il confronto delle diverse strategie di trattamento, sotto l'assunzione di non proporzionalità delle funzioni di sopravvivenza. Questo metodo risulta particolarmenteutile quando le funzionidi sopravvivenza stimata si incrociano tra loro. Gli studi di simulazione condotti su questo metodo hanno mostrato che esso presenta una potenza più elevata rispetto al test pesato dei ranghi logaritmici, nel caso in cui le curve di sopravvivenza si incrociano e non sono quindi proporzionali tra loro. L'ultima parte della tesi si concentra sull'analisi di eventi avversi nell'ambito degli studi randomizzati a due stadi. È stata sviluppata una metodologia per analizzare dati relativi ad eventi avversi, che si basa anche sui modelli a rischi competitivi. Questa metodologia è stata poi applicata per analizzare dati di eventi avversi in prove cliniche di trattamenti per la leucemia.

Acknowledgements

I would like to thank my supervisor, Professor Giuliana Cortese, for her guidance throughout this PhD project. Without her support this work would not have been possible. I would also like to thank the Alliance Concepts for providing me with the dataset I used in this thesis. Many thanks to Dr. Sumithra Mandrekar from the Mayo Clinic, Morgen Alexander-Young from the Alliance Concepts and Prof. Richard A. Larson for their assistance in obtaining the dataset.

Finally, I would like to thank my family members for supporting me in this three years.

Contents

List of Figures

List of Tables

Introduction

Overview

Treatment and management of chronic illnesses such as cancer, leukemia and HIV often require multiple courses of treatment. The clinical goal in such trials is to achieve a response such as complete remission of leukemia, 50% shrinkage of solid tumor or increase in CD4 count in HIV patients. These responses are presumed to predict longer survival. Dynamic treatment regimes, also known as dynamic treatment strategies or treatment policies, have become popular in the conduct of cancer trials (Lokhnygina and Helterbrand, 2007). These designs use a sequence of decision rules that link the observed patient's history with treatment recommendations. In two-stage randomization designs, for instance in cancer clinical trials, patients are initially randomized to an induction treatment followed by another randomization into a maintenance regimen provided that the patient responds to the induction therapy and consents to further study. These designs are sometimes referred to as sequential multiple assignment randomized trials (SMART).

The focus in two-stage randomization designs with survival endpoints is on estimating survival distributions and comparing the different treatment policies. Several methods for estimating the survival distributions have been proposed in the literature. The weighted risk set estimator (WRSE) is an extension of the Nelson-Aalen estimator by incorporating inverse probability weights (Guo and Tsiatis, 2005). This estimator is derived based on the counting processes. Another estimator was proposed by Lunceford et al. (2002) which also uses inverse probability weights. Miyahara and Wahed (2010) proposed the weighted Kaplan-Meier estimator for the estimation of survival distributions in two-stage randomization designs. This estimator is a modification of the usual Kaplan-Meier estimator. Wahed (2010) developed a parametric inference approach based on mixture distributions to compare different treatment policies. The comparison is based on the survival means for the treatment strategies. In the same

paper, Wahed (2010) also proposed a method for estimating survival distributions for two-stage randomization designs where response to the first stage treatment is measured at a single fixed time point. SMART design where response is measured at a fixed point in the first stage are known as standard SMART designs. In time-varying SMART designs, patients are randomized to second stage as soon as a response to the first stage treatment is observed.

In his unpublished thesis, Guo (2005) proposed a weighted version of the log-rank test for comparing separate-path treatment policies. Treatment policies are regarded as separate-path if they do not share the same first stage treatment. Kidwell and Wahed (2013) extended Guo (2005) approach to include shared path treatment strategies. Two treatment strategies are shared-path if they share the same first stage treatment. However, the paper by Kidwell and Wahed (2013) did not directly addressed the case where the survival curves cross. In this thesis, we propose a testing approach based on the absolute difference of the area under the survival curves which works well even if the survival curves cross.

In this thesis, we review some non-parametric methods used in the estimation of survival distributions for two-stage randomization designs. We provide a simulation study to compare these methods when extreme values of the simulation parameters are used. A parametric approach for estimating survival distributions for time-varying SMART designs is introduced. The method is evaluated using a simulation study and also applied to a dataset from a clinical trial. Safety data are not usually given the same attention as efficacy data in the analysis of clinical trials data (Allignol et al., 2016). In this thesis we develop a methodology for analyzing safety data from two-stage randomization designs.

Main contributions of the thesis

The main contributions of this thesis can be summarized as follows,

- Comparative study of the non-parametric estimators for estimating survival distributions in two-stage randomization designs.
- Proposal of a parametric method for estimating survival distributions for time-varying SMART designs. Numerical evaluation of the proposed estimator. Application to the CALGB 19808 study dataset. This dataset is from a leukemia clinical trial for patients under the age of 60 obtained from Mayo Clinic (Kolitz et al., 2010).

- Introduction of the weighted Lin and Xu test for comparing treatment policies. Extension of the Lin and Xun test to accommodate shared-path treatment policies.
- Development of competing risk model and the related inference for analyzing safety data from two-stage randomized designs. Application to the CALGB study 19808 toxicity dataset.

Chapter 1

A review of methods for estimating survival distributions

Randomized controlled trials (RCTs) are regarded as the gold standard in biomedical research. When they can be implemented, RCTs provide the strongest evidence in terms of the unbiasedness and consistency of the estimate of the treatment effect. Though regarded as the gold standard, there are criticisms against RCTs. One of these criticisms is the one size fits all approach. To remedy this, there has been a shift from the traditional one size fits all approach. Researchers in many fields are acknowledging that single interventions may be limited in their effectiveness due to heterogeneity on treatment response between individuals or within an individual over time. Even in real life situations, individuals use multiple interventions. For example, if weight loss is the objective, an individual may increase from exercising once per week to say three times a week. If exercise does not yield the desired outcome, the individual may switch to another intervention such as dieting or to a combination of both. Switching interventions or sequences of interventions as opposed to a single stage intervention can have a huge effect on the outcome.

1.1 Dynamic treatment regimes

Adaptive interventions, also known as dynamic treatment regimes use a sequence of decision rules which recommend when and how the intervention should be modified in order to optimize long term primary outcomes. These recommendations are based on factors such as individual characteristics, intermediate response collected in the course of the intervention such as the individual's response and adherence. In adaptive treatment strategies, the intervention is personalized depending on the specific needs of the

individual, secondly, the intervention is time varying, that is, the intervention is repeatedly adapted overtime in response to the participant's performance and changing needs. An important aspect of adaptive interventions is the periodic assessment to ascertain weather the treatment selected initially is in fact helpful. If not helpful, adaptation procedures become necessary and sometimes these adjustments are done several times during the course of the treatment (Kidwell and Hyde, 2016).

One way to operationalize adaptive interventions is to use decision rules to link individual's characteristics and ongoing performance with specific treatment options. Adjustments of intervention options are based on the individual's values on tailoring variables. Candidates for tailoring variables depends largely on the problem at hand. An individual's responsiveness to an intervention is considered as an important tailoring variable in many clinical trials. Response is defined based on the disease under study. For an example, in an HIV clinical trial, reaching a certain threshold in CD4 count may be regarded as a response. Alternatively, the choice of intervention options can also be tailored based on treatments already received.

A multistage randomized trial in which each stage corresponds to a decision is referred to as a SMART design. At each stage of the trial, individuals are assigned to one of the several treatment options. Data from SMART designs can be used in addressing questions concerning comparison of different interventions, they can also be used in making comparisons of different treatment regimes embedded in the SMART design (Nahum-Shani et al., 2012).

1.1.1 Time-varying SMART designs

We differentiate between two SMART designs. SMART designs with outcome assessments at fixed time point are referred to as standard SMART designs (Dai and Shete, 2016). In such a SMART design, we take the time to response to be the same for every individual in the study. This is because the time to measure response to the first stage treatments is fixed for every individual, for example at six months. Figure 1.1 shows a standard SMART design where individuals in the first stage are randomized to either A_1 or A_2. Responders are randomized to either B_1 or B_2 in the second stage.

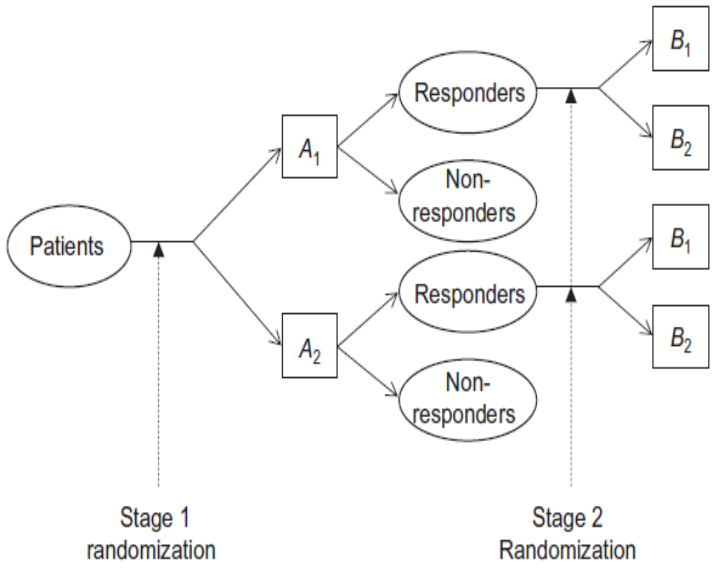


FIGURE 1.1: An example of a two-stage standard SMART design.

In a time-varying SMART design, individuals are randomized to the second stage treatments as soon as a response is observed. This means that the response times for the individuals vary. There are, however, advantages for this type of design especially in cancer trials where medications have some side effects. Prolonged intake of such medications in the first stage even when a response has been observed may lead to several side effects and thereby making patients refused second stage treatments. Also this type of design may lead to reduction in costs (Dai and Shete, 2016). Figure 1.2 shows an example of a time-varying SMART design where only responders are randomized to the second stage. In this figure, individual 3 responds at time T_3^R and is then randomized to the second stage treatments. We do not show the second stage randomization for the other individuals to avoid having many lines on the figure.

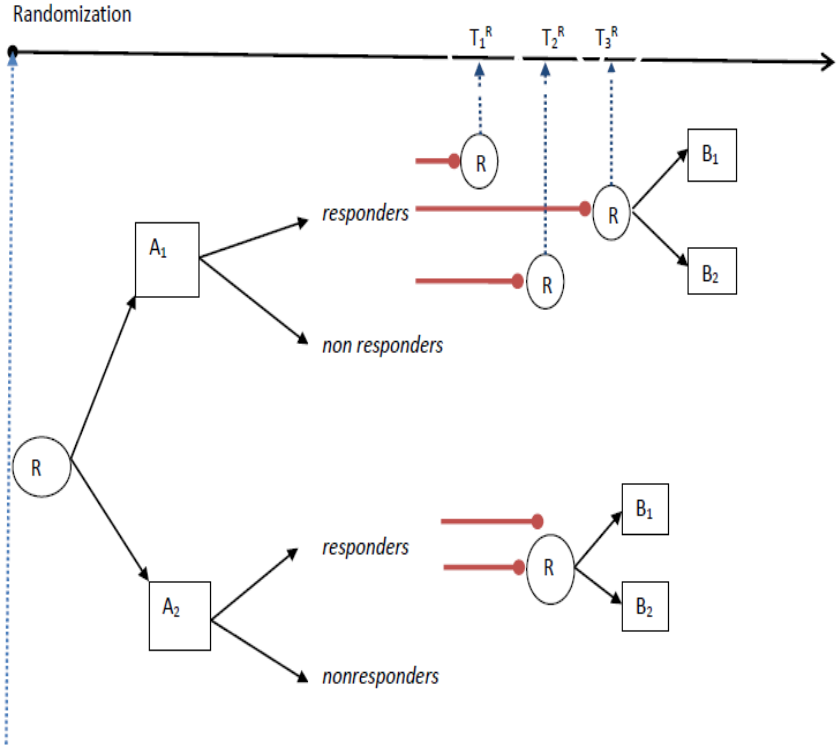


FIGURE 1.2: An example of a two-stage time-varying SMART design.

1.2 Counterfactuals

Counterfactuals play an important role in causal inference. The focus in causal inference is on the effects of causes or treatments on specific units. In a clinical trial the units are the individuals in the study. Consider a study where half of the units are exposed to the experimental treatment (E) and the other half exposed to the control treatment (C). If the treatments E and C are assigned randomly to the experimental units such that each unit was equally likely to be exposed to E as to C, then such a study is referred to as a randomized experiment, otherwise the study is an observational study or a non-randomized study. For a particular unit, say u, the causal effect of E versus C is the difference between what would have happened at some time, say t_2, if u had been exposed to E initiated at t_1, and what would have happened if u was exposed to C at t_2 started at t_1. Denote by Y the outcome variable such that $Y_E(u)$ and $Y_C(u)$ are the outcomes if unit u is assigned to E or C, respectively.

The effect of the treatment E on the unit u as measured by Y at t_2, and relative to treatment C, is the difference $Y_E(u) - Y_C(u)$. Defining the causal effect in this manner leads to a problem. Practically it is impossible to observe the value of $Y_E(u)$ and $Y_C(u)$ on the same unit u. $Y_E(u)$ and $Y_C(u)$ are called potential outcomes or

counterfactuals. From unit u, either $Y_E(u)$ or $Y_C(u)$ is observed but not both, which means that one of these potential outcomes is always missing. From a statistical view point, the estimation of causal effects can be regarded as a particular problem of missing data. One of the important assumptions in causal inference is the stable unit treatment assumption which states that what is observed in one unit should not be affected by the particular assignment of treatments to other units (Holland, 1986).

1.3 Model framework

Consider a two-stage randomization design where patients are first randomized to receive treatment A with levels A_1 and A_2, and individuals who respond and consent to further study are randomized to the second treatment with levels say B_1 and B_2. For simplicity we shall henceforth use the word "response" to indicate response to previous treatment and consent to second randomization. The strategy A_jB_k, for $j, k = 1, 2$ entails treating with A_j followed by B_k if the patient responds to the first treatment. Our objective is to estimate and compare survival distributions for the different treatment policies. For this scope, we conceptualize the problem using potential outcomes (Rubin, 1974). This does not mean that focus is on causal inference but we use potential outcomes as a vehicle for formalizing the problem.

In reality, each individual follows only one treatment strategy, we observe only one outcome for the specific treatment strategy. However, in theory individuals in the population can follow any treatment policy A_jB_k, that is, for each individual in the population one can envision one outcome for each possible strategy. Each individual has his/her set of potential outcomes, the entire set of possible outcomes for each individual is referred to as his/her counterfactuals.

Here, we shall focus on data from patients who received induction therapy A_1, since data from patients who received different induction therapies are independent. Data from patients who received A_2 are analyzed in a similar manner. Interest is on estimating survival distributions for treatment policies A_1B_1 and A_1B_2. We assume that, associated with subject i is a set of potential outcomes

$$\left\{R_i^*, (1 - R_i^*)T_{0i}, R_i^*T_i^R, R_i^*T_{1i}^*, R_i^*T_{2i}^*\right\}$$

where R_i^* is the response status if patient i was assigned to A_1. $R_i^* = 1$ if patient i responds to treatment A_1, $R_i^* = 0$ otherwise. T_i^R is the time from initial randomization to response for patient i defined only when $R_i^* = 1$; T_{0i} is the survival time for a patient

who do not respond to first stage treatment. T^*_{1i} is the time from second randomization to death if patient i receives B_1, and similarly T^*_{2i} is the time from second randomization to death if patient i receives B_2 instead. If patient i is assigned to A_1B_k, his/her survival time would be

$$T_{ki} = (1 - R^*_i)T_{0i} + R^*_i(T^R_i + T^*_{ki}), \quad k = 1, 2. \tag{1.1}$$

We note that we can only observe T_{1i} or T_{2i} hence, T_{ki} are potential outcomes. If $R^*_i = 0$ then $T_{1i} = T_{2i} = T_{0i}$.

Let T_k denote the survival time for the population if all participants were assigned to the treatment strategy A_1B_k. Inference on features of these distributions address directly the intent-to-treat question of interest. Using data from two-stage design we estimate the distribution of T_k.

Without right censoring, the observed data can be represented as a set of independent and identically distributed (iid) random vectors $(R^*_i, R^*_iT^R_i, R^*_iZ_i, T_i)$ for $i = 1, \ldots, n$, where Z_i is an indicator for the B treatment defined only if $R^*_i = 1$. We have $Z_i = 1$ if patient i is assigned to B_1 and $Z_i = 0$ if assigned to B_2. The observed survival time, T_i, is related to the potential outcomes as follows:

$$T_i = (1 - R^*_i)T_{0i} + R^*_i\left\{T^R_i + Z_iT^*_{1i} + (1 - Z_i)T^*_{2i}\right\}. \tag{1.2}$$

To incorporate right censoring, let C_i be the time to censoring for the ith patient. The observed data can then be represented as independent and identically distributed vectors $(R_i, R_iZ_i, R_iT^R_i, U_i, \Delta_i)$, where $\Delta_i = I(T_i < C_i)$ is the failure indicator, $U_i = \min(T_i, C_i)$ is the observed time to either death or censoring. $R_i = 0$ if patient i is censored without having had a response to treatment A_1 otherwise $R_i = R^*_i$.

We assume that the second stage randomization is made independently of the other potential outcomes, that is

$$\begin{aligned}\pi_z &= P(Z_i = 1|R_i = 1, T^R_i, T_{1i}, T_{2i}, C_i)\\ &= P(Z_i = 1|R_i = 1).\end{aligned}$$

We note that π_z, defined only if $R_i = 1$, is the probability of being randomized to the B treatment and it is typically fixed by design.

1.4 Standard methods for estimating survival distributions

Prior to the introduction of the weighted methods for analyzing data from two-stage randomization designs, the analysis was typically separated into two parts under an intent-to-treat principle, where patients are analyzed according to the group to which they were randomized. The data from the first stage was analyzed separately ignoring the second stage data. A similar analysis was done for the second stage data. With time-to-event endpoints, standard survival analysis methods are used to analyze the data from the two stages. We give a brief overview of the standard methods for estimating survival distributions.

1.4.1 Kaplan-Meier estimator

The widely used estimator for the survival function is the Kaplan-Meier estimator. Let $t_1 < t_2 < ...$ be the distinct event times, suppose that at time t_m for $m = 1, 2, ...$ there are d_m events. Let Y_m be the number of individuals at risk at time t_m, that is, the set of individuals who did not yet have the event until time $t_m{}^-$, where $t_m{}^-$ is the time instant just before t_m. The Kaplan-Meier estimator is as follows:

$$\hat{S}(t) = \begin{cases} 1, & t < t_1 \\ \prod_{t_m \leq t} \left(1 - \frac{d_m}{Y_m}\right), & t \geq t_1. \end{cases}$$

The Kaplan-Meier estimator is a step function with jumps at the observed event times. The variance of the Kaplan-Meier estimator is estimated using the Greenwood's formula.

1.4.2 Nelson-Aalen estimator

One possible estimator for the cumulative hazard, $\Lambda(t) = \int_0^t \alpha(s)ds$, is the Nelson-Aalen estimator, defined as

$$\hat{\Lambda}(t) = \begin{cases} 0, & t \leq t_1 \\ \sum_{t_m \leq t} \frac{d_m}{Y_m}, & t \leq t_1. \end{cases}$$

The variance of the Nelson-Aalen estimator is estimated by

$$\hat{V}(\hat{\Lambda}(t)) = \sum_{t_m \leq t} \frac{d_m}{Y_m^2}.$$

The Nelson-Meier estimator for the cumulative hazard rate can be used as an alternative estimator for the survival function since $\hat{S}(t) = \exp(-\hat{\Lambda}(t))$.

The use of standard survival methods cannot answer questions regarding treatment sequences. One cannot tell which treatment policy leads to longer survival and also the data is not used efficiently if analysis is done separately for the two stages. To answer questions on treatment policies, an alternative design would randomize the patients up front to the treatment policies. This approach, however, requires a large number of patients and it is not efficient (Lunceford et al., 2002).

1.5 Estimating survival distributions for two-stage designs

Several nonparametric estimators have been proposed. The most popular ones are the weighted risk set estimator (WRSE) of Guo and Tsiatis (2005), the inversely weighted estimators proposed by Lunceford et al. (2002) which we shall refer to as the LDT estimator and the weighted Kaplan-Meier estimator of Miyahara and Wahed (2010).

1.5.1 LDT estimator

The LDT estimator (Lunceford et al., 2002) is derived using the inverse weighting technique (Robins et al., 1994). Consider the estimation of the survival distributions for the treatment policy A_1B_k, that is, $S_{1k}(t) = 1 - P(T_{1k} \leq t) = 1 - F_{1k}$, for $k = 1, 2$. For simplicity, consider A_1B_1. In two-stage designs, difficulties arise from subjects who are not consistent with the treatment policy of interest. In this case we treat them as missing. If all the patients are assigned to A_1B_1 and there is no censoring, meaning $U_i = T_i = T_{1i}$, the natural estimator for $F_{11}(t)$ is $n^{-1}\sum_{i=1}^{n} I(U_i \leq t)$. With censoring and second stage randomization upon response, only a subset of patients would have their observed survival time and actual treatment received being consistent with A_1B_1 since some patients are randomized to A_1B_2. Lunceford et al. (2002) proposed an estimator based on inverse probability weighting (Robins et al., 1994) to weight observations in this subset in such a way that the distribution of the weighted observations mimic that in an ideal case. Let $W_{1i} = 1 - R_i + R_iZ_i/\pi_z$ be the weight function. When the ith patient is consistent with treatment policy A_1B_1, W_{1i} acts as a weight. Non-responders consistent with A_1B_1 represent themselves and they get a weight of 1, that is, $W_{1i} = 1$. Each responder consistent with A_1B_1 represents $1/\pi_z$ remitting or consenting individuals who could have

been potentially assigned to B_1 and gets a weight of $1/\pi_z$. Responders not consistent with the policy A_1B_1 get a weight of 0.

This weighting scheme motivates the estimator

$$\hat{F}^1_{1k}(t) = n^{-1} \sum_{i=1}^{n} \frac{\Delta_i W_{ki}}{\hat{K}(U_i)} I(U_i \leq t), \quad k = 1, 2; \tag{1.3}$$

where $\hat{K}(U_i)$ is the Kaplan-Meier estimator for the censoring distribution given by $\hat{K}(U_i) = \prod_{u \leq t}\{1 - dN^c(u)/Y(u)\}$ with $N^c = \sum_{i=1}^{n} I(U_i \leq u, \Delta_i = 0)$ and $Y(u) = \sum_{i=1}^{n} I(U_i \geq u)$.

Instead of dividing by n in (1.3), a second estimator can be obtained by dividing by a probabilistically adjusted sample size;

$$\hat{F}^*_{1k}(t) = \left\{ \sum_{i=1}^{n} \frac{\Delta_i W_{ki}}{\hat{K}(U_i)} \right\}^{-1} \sum_{i=1}^{n} \frac{\Delta_i W_{ki}}{\hat{K}(U_i)} I(U_i \leq t), \quad k = 1, 2. \tag{1.4}$$

From (1.4), the survival distributions for A_1B_k are estimated using

$$\hat{S}_{1k}(t) = 1 - \hat{F}^*_{1k}(t); \tag{1.5}$$

and the variance is estimated by

$$\widehat{var}(\hat{S}_{1k}(t)) = \frac{1}{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i W_{ki}}{\hat{K}(U_i)} \{I(U_i \leq t) - \hat{F}^*_{1k}\}^2 + \int_0^L \frac{dN^c(u)}{\hat{K}(u)Y(u)} \hat{E}\{L^*_{1ki}(t,u)\}^2 \right\}, \tag{1.6}$$

where L is the restricted lifetime which is smaller than the maximum follow-up of the study,

$$E\{L^*_{1ki}(t,u)\}^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(U_i)} \left[W_{ki}\{I(U_i \leq t) - \hat{F}^*_{1k}(t)\} - \hat{G}^*_{1k}(t,u) \right]^2 I(U_i \geq u), \tag{1.7}$$

and

$$\hat{G}^*_{1k}(t,u) = \{n\hat{S}(u)\}^{-1} \sum_{i=1}^{n} \frac{\Delta_i W_{ki}}{\hat{K}(U_i)} \left\{ \{I(U_i \leq t) - \hat{F}^*_{1k}(t)\} \right\} I(U_i \geq u). \tag{1.8}$$

More details on the variance derivation can be found on the appendix of Lunceford et al. (2002).

1.5.2 Weighted risk set estimator

The derivation of the WRSE estimator relies heavily on the counting processes. For a one-stage study with survival endpoints, the cumulative hazard rate can be estimated by the Aalen-Nelson estimator

$$\hat{\Lambda}(t) = \int_0^t \frac{dN(u)}{Y(u)},$$

where $N(u) = \sum_{i=1}^{n} I(U_i \leq u, \Delta_i = 1)$ denotes the number of deaths up to and including time u, and $Y(u) = \sum_{i=1}^{n} I(U_i \geq u)$ is the number of patients at risk at time u. The WRSE is here showed for A_1B_1, as the development of the estimator for A_1B_2 follows similarly. Consider the case when all individuals are assigned to A_1B_1 in which case the observed death or censoring time is $U_{1i} = \min(T_{1i}, C_i)$. Let $N_{1i}(u) = I(U_{1i} \leq u, \Delta_i = 1)$ and $Y_{1i}(u) = I(U_{1i} \geq u)$ then the cumulative hazard estimator becomes

$$\hat{\Lambda}_{11}(t) = \int_0^t \frac{\sum_{i=1}^{n} dN_{1i}(u)}{\sum_{i=1}^{n} Y_{1i}(u)}.$$

In reality, some of the patients who could have received B_1 received instead B_2 after randomization to the second stage. $N_{1i}(u)$ and $Y_{1i}(u)$ cannot be observed directly and the WRSE propose to incorporate inverse weighting where the weight function depending on u is defined as $W_i(u) = 1 - R_i(u) + R_i(u)Z_i/\pi_z$, where $R_i(u)$ is the response status at time u. $R_i(u) = 0$, if at time u a response has not been achieved for patient i but patient i is still consistent with A_1B_1 and gets a weight of 1. For a patient i with $R_i(u) = 1$ and $Z_i = 0$, a weight of 0 is assigned since this patient is no longer consistent with the treatment strategy A_1B_1. For a responder assigned to B_1, this patient is consistent with A_1B_1 and gets a weight of $1/\pi_z$ at time u. This patient represents $1/\pi_z$ individuals who could have been potentially assigned to B_1. The weight function $W_i^*(u) = 1 - R_i(u) + R_i(u)(1 - Z_i)/(1 - \pi_z)$ is used for A_1B_2 and a similar argument is made.

The difference between the LDT and the WRSE is that the WRSE uses time dependent weights. A patient who is a responder and is randomized to B_2 gets a weight of 0 under the LDT at any time u including the time before the second randomization. This leads to a loss in efficiency. On the contrary, the WRSE includes this subset of patients and assigns a weight $W_i = 1$ at any time u before the second randomization, thereafter the weight changes to $W_i = 0$.

The cumulative hazard estimator for A_1B_1 using the above weight function is

$$\hat{\Lambda}_{11}(t) = \int_0^t \frac{\sum_{i=1}^n W_i(u)dN_i(u)}{\sum_{i=1}^n W_i(u)Y_i(u)}$$

where $N_i(u) = I(U_i \leq u, \Delta_i = 1)$ and $Y_i(u) = I(U_i \geq u)$. The survival estimator is

$$\hat{S}_1(t) = \exp\left\{-\int_0^t \frac{\sum_{i=1}^n W_i(u)dN_i(u)}{\sum_{i=1}^n W_i(u)Y_i(u)}\right\}. \tag{1.9}$$

The variance is given by

$$\widehat{\text{Var}}(S_{A_1B_1}(t)) = n^{-1}\{S_{A_1B_1}(t)\}^2\hat{\sigma}^2, \tag{1.10}$$

where

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n \left(\int_0^t \frac{W_i(u)\left[dN_i(u) - Y_i(u)\left\{\frac{\sum_{i=1}^n W_i(u)dN_i(u)}{\sum_{i=1}^n W_i(u)Y_i(u)}\right\}\right]}{n^{-1}\sum_{i=1}^n W_i(u)Y_i(u)}\right)^2$$

1.5.3 Weighted Kaplan-Meier estimator

Miyahara and Wahed (2010) proposed two forms of the weighted Kaplan-Meier estimators for two-stage randomization designs. Consider the case where all patients are treated with A_1B_k, the survival function at time t can be estimated as

$$\hat{S}_{A_1B_k}(t) = \begin{cases} 1, & t < t_1 \\ \prod_{t_m \leq t}\left(1 - \frac{d_m}{Y_m}\right), & t \geq t_1, \end{cases}$$

for $k = 1, 2$, where $d_m = \sum_{i=1}^n \Delta_i I(U_i = t_m)$, $Y_m = \sum_{i=1}^n I(U_i \geq t_m)$ and t_m are the ordered death or failure times for $m = 1, 2, \ldots$. However, we know that in a two-stage randomization design some patients will potentially receive treatment not consistent with A_1B_k, so an adjustment for such a loss is done using inverse probability weighting. The weighted Kaplan-Meier estimator (WKME) is

$$\hat{S}^w_{A_1B_k}(t) = \begin{cases} 1, & t < t_1 \\ \prod_{t_m \leq t}\left(1 - \frac{d^w_m}{Y^w_m}\right), & t \geq t_1, \end{cases} \tag{1.11}$$

for $k = 1, 2$ and where $d^w_m = \sum_{i=1}^n \Delta_i I(U_i = t_m)W_{ki}$, $Y^w_m = \sum_{i=1}^n I(U_i \geq t_m)W_{ki}$. The death process and the at risk process are weighted by the inverse probability of randomization. The weights used in this estimator are the same as in the LDT estimator

above. A modified version of the Greenwood formula can be used to estimate the variance of the estimator

$$\widehat{var}(\hat{S}_{A_1B_k}(t)) = \{\hat{S}_{1k}(t)\}^2 \sum_{m:t_m \leq t} \frac{1 - \hat{s}_{1km}}{\hat{M}_{1km}\hat{s}_{1km}}, \tag{1.12}$$

where

$$\hat{M}_{1km} = \left(\sum_{i=1}^{n} W_{ki} I(U_i \geq t_m) \right)^2 / \sum_{i=1}^{n} \{W_{ki} I(U_i \geq t_m)\}^2, \tag{1.13}$$

and $\hat{s}_{1km} = 1 - d_m^w / Y_m^w, m = 1, 2, ...$

A version of this estimator with time-dependent weights was also developed. For the estimator with time dependent weights, the variance was calculated using bootstrap (Efron and Tibshirani, 1994)

1.6 Sensitivity analysis

We performed a simulation study to compare the performance of the three non-parametric methods, namely the WRSE, LDT and the WKM estimators. The aim of this simulation study is to ascertain how these methods perform when extreme values of the parameters are used. We note that in the simulation studies on WKM estimators reported in the paper by Miyahara and Wahed (2010), only two response rates were used (0.4 and 0.7) and also, in their simulation scenarios, the censoring rates were 5.4% for the 40% response rate and 6.4% for the 70% response rate. The papers on the LDT estimator (Lunceford et al., 2002) and the WRSE (Guo and Tsiatis, 2005) used only one censoring rate in their simulations. In the paper about the WRSE (Guo and Tsiatis, 2005), two response rates were used, (0.5 and 0.8). In the LDT estimator simulations, 20% and 50% were used as response rates. In real world application, one can find datasets with higher censoring rates than those considered in the simulation studies of the above papers, and perhaps higher or relatively lower response rates than the ones used in these studies. How do these methods perform when extreme values of the simulation parameters are used? How are these methods affected by higher censoring rates?

In this simulation study, we used the following censoring rates: 0%, 10%, 30%, 50% and 60%. For the response rates we used 20%, 40%, 60%, and 80%. We considered three sample sizes, that is, $n = 100, 200, 400$. The simulation study is done similar to that of the WRSE. The response and consent indicator, R_i, is simulated from a Bernoulli

distribution with $P(R_i = \theta)$, where $\theta = (0.2, 0.4, 0.6, 0.8)$. For non-responders we simulated T_{0i} from an exponential distribution with mean 182.5 days. T_{0i} is generated only if $R_i = 0$. The time to response, T_i^R, is generated following an exponential distribution with mean 300 days. In all the simulation scenarios, the indicator for the B-treatment is generated from a Bernoulli (0.5) distribution. If $Z = 1$, T_{1i}^* is generated from an exponential distribution with mean 370 days and if $Z = 0$, T_{2i}^* is generated from an exponential distribution with mean 547.5 days.

The observed survival time for patient i is defined as $T_i = (1-R_i)T_{0i}+R_i\{T_i^R+Z_iT_{1i}^*+(1-Z_i)T_{2i}^*)$. The censoring time was generated from a uniform $(0, v)$ distribution, v is chosen such that 0%, 10%, 30%, 50% and 60% of the times are censored. The observed survival time is defined to be $U_i = \min(T_i, C_i)$. For simplicity, we only calculate survival distributions for A_1B_1.

In all the simulation scenarios, 1000 datasets were generated and the methods were applied. We estimated $S_1(t)$ at $t = 100, 300, 450$ days. We report the survival probabilities together with their standard errors, coverage probabilities, and bias for the three methods at the times mentioned above. In addition, we also report the relative efficiency (RE) of WRSE and WKM estimators which is calculated as $RE =$ sample variance of WRSE/sample variance of WKM. Guo and Tsiatis (2005) established that the WRSE is more efficient than the LDT estimator, so we did not repeat the calculation of the relative efficiency here.

We present the results of the simulation study for $n = 100$ in Table 1.1, $n = 200$ in Table 1.3 and $n = 400$ in Table 1.3. The LDT estimator is mostly affected by the censoring rates. In all the sample sizes, the bias of the LDT estimator increases drastically as the censoring rate is increased independently of the response rate, (π_r). At the lower end of the survival curve, the LDT estimator approaches zero in cases where the response rate is low and the censoring rate is high. This is more evident when the sample size is small, that is, $n = 100$. The calculation of of the LDT depends on the censoring distribution. Higher values of the censoring distribution leads to a bigger denominator, hence this estimator approaches zero at the lower end when high censoring rates are used. For low censoring rates, all three estimators yield similar estimates, this happens when the censoring rate is less than 30%. The bias of the LDT estimator becomes greater when the censoring rate is 30% and above. The bias of the WRSE and the WKM estimator also increase when the censoring rate is high, and becomes more pronounced in the lower end of the survival curve.

The LDT estimator has the best coverage probabilities in cases where the censoring rate is low. In cases where the censoring rate is high, the coverage probabilities for the

LDT estimator are far from the nominal level of 95%. The coverage probabilities of the WRSE and the WKPE are are also affected by high censoring rates and low response rates. For a sample size of $n = 100$, and response rate of 20%, the coverage probabilities are 47.4% and 50.6% for the WRSE and the WKM estimator for the censoring rate of 60%. With increase in response rates, the coverage probabilities for the WRSE and the WKM estimator get closer to the desired nominal level. For a response rate of 80% and a censoring rate of 60%, the coverage probabilities for the WRSE and the WKPM estimator are 93.0% and 85.3% respectively. Increasing the sample size from $n = 100$ to $n = 400$, the coverage probabilities get closer to the desired nominal level for the WRSE and the WKM estimator whilst the coverage probabilities of the LDT estimator lags behind when the censoring rate is high. At the lower end of the survival curve, the coverage probabilities for the WKM estimator are not close to the nominal level.

As expected, increasing the sample size leads to decrease in the standard errors for the three methods. In all the simulation scenarios, the LDT estimator has the largest standard errors. This is a result that was established in the Guo and Tsiatis (2005) paper. The standard errors for the WRSE and the WKM estimator are similar. There is a small gain in efficiency in using the WKM estimator. The WRSE and the WKM estimator are less affected by high censoring rates and low response rates. The survival estimates from the WRSE and the WKM estimator are similar and their standard errors are similar as well.

TABLE 1.1: Simulation results for $n = 100$

		WRSE				**WKM**				**LDT**			
t	$S_1(t)$	$\hat{S}_1(t)$	SE	Bias	CP	$\hat{S}_1(t)$	SE	Bias	RE	CP	$\hat{S}_1(t)$	SE	Bias	CP
							$\pi_r = 0.2$	$c = 0\%$						
100	0.655	0.655	0.047	0.00	93.9	0.653	0.047	0.00	1.01	93.2	0.653	0.050	0.00	94.9
300	0.309	0.311	0.048	0.00	92.4	0.308	0.046	0.00	1.11	93.0	0.307	0.054	0.00	95.0
450	0.190	0.192	0.042	0.00	92.4	0.188	0.038	0.00	1.20	88.6	0.187	0.048	0.00	93.3
							$\pi_r = 0.2$	$c = 10\%$						
100	0.655	0.656	0.048	0.00	93.2	0.654	0.048	0.00	1.01	92.9	0.654	0.051	0.00	93.9
300	0.309	0.312	0.050	0.00	94.2	0.308	0.047	0.00	1.10	93.0	0.307	0.058	0.00	94.2
450	0.190	0.198	0.045	0.01	95.1	0.192	0.041	0.00	1.21	88.7	0.190	0.053	0.00	94.4
							$\pi_r = 0.2$	$c = 30\%$						
100	0.655	0.656	0.049	0.00	95.4	0.653	0.049	0.00	1.00	94.6	0.626	0.052	0.03	88.9
300	0.309	0.313	0.054	0.00	93.2	0.307	0.051	0.00	1.10	94.1	0.253	0.057	0.06	70.9
450	0.190	0.196	0.051	0.01	91.8	0.186	0.047	0.00	1.17	89.9	0.124	0.047	0.07	56.9
							$\pi_r = 0.2$	$c = 50\%$						
100	0.655	0.658	0.051	0.00	95.7	0.654	0.051	0.00	1.00	94.8	0.547	0.050	0.10	46.9
300	0.309	0.318	0.066	0.00	93.1	0.305	0.065	0.00	1.02	91.0	0.098	0.035	0.21	16.1
450	0.190	0.247	0.066	0.05	75.5	0.222	0.067	0.03	0.87	74.2	0.000	0.000	0.19	0.00
							$\pi_r = 0.2$	$c = 60\%$						
100	0.655	0.659	0.053	0.00	94.3	0.657	0.053	0.00	1.00	93.7	0.481	0.049	0.17	22.3
300	0.309	0.343	0.076	0.03	83.6	0.322	0.078	0.01	0.89	81.8	0.000	0.000	0.30	0.00
450	0.190	0.343	0.076	0.15	47.4	0.322	0.078	0.13	0.89	50.6	0.000	0.000	0.19	0.00
							$\pi_r = 0.4$	$c = 0\%$						
100	0.732	0.732	0.045	0.00	93.5	0.730	0.044	0.00	1.03	92.3	0.731	0.048	0.00	95.2
300	0.425	0.428	0.053	0.00	94.6	0.423	0.049	0.00	1.18	92.0	0.425	0.060	0.00	95.1
450	0.295	0.298	0.051	0.00	95.3	0.293	0.045	0.00	1.29	88.7	0.295	0.059	0.00	95.9
							$\pi_r = 0.4$	$c = 10\%$						
100	0.732	0.734	0.045	0.00	93.7	0.731	0.044	0.00	1.04	91.5	0.732	0.049	0.00	95.0
300	0.425	0.429	0.055	0.00	94.1	0.424	0.050	0.00	1.18	91.6	0.425	0.063	0.00	94.7
450	0.295	0.300	0.053	0.00	93.9	0.295	0.047	0.00	1.29	87.6	0.295	0.062	0.00	94.3
							$\pi_r = 0.4$	$c = 30\%$						
100	0.732	0.734	0.046	0.00	92.6	0.731	0.045	0.00	1.03	93.7	0.714	0.050	0.01	91.0
300	0.425	0.431	0.057	0.01	93.7	0.425	0.053	0.00	1.17	92.1	0.386	0.067	0.04	83.5
450	0.295	0.304	0.057	0.01	92.8	0.295	0.051	0.00	1.27	88.3	0.248	0.067	0.05	76.6
							$\pi_r = 0.4$	$c = 50\%$						
100	0.732	0.735	0.047	0.00	94.5	0.732	0.046	0.00	1.04	91.6	0.640	0.050	0.09	53.3
300	0.425	0.430	0.065	0.01	91.9	0.422	0.060	0.00	1.16	92.1	0.228	0.058	0.19	26.6
450	0.295	0.305	0.072	0.01	88.5	0.289	0.067	0.01	1.14	82.3	0.057	0.023	0.24	14.7
							$\pi_r = 0.4$	$c = 60\%$						
100	0.732	0.735	0.048	0.00	93.7	0.732	0.047	0.00	1.03	91.9	0.584	0.049	0.14	29.7
300	0.425	0.431	0.072	0.01	93.0	0.420	0.069	0.01	1.11	87.7	0.110	0.036	0.31	9.90
450	0.295	0.360	0.078	0.06	74.3	0.335	0.074	0.04	1.00	71.4	0.000	0.000	0.29	0.00
							$\pi_r = 0.6$	$c = 0\%$						
100	0.809	0.809	0.041	0.00	94.2	0.806	0.039	0.00	1.08	90.9	0.807	0.044	0.00	94.7
300	0.541	0.543	0.056	0.00	93.4	0.538	0.050	0.00	1.27	91.7	0.540	0.062	0.00	95.0
450	0.400	0.406	0.057	0.01	93.8	0.399	0.049	0.00	1.38	88.7	0.401	0.064	0.00	95.5
							$\pi_r = 0.6$	$c = 10\%$						
100	0.809	0.811	0.041	0.00	93.2	0.810	0.039	0.00	1.09	90.7	0.810	0.044	0.00	93.6
300	0.541	0.548	0.057	0.01	94.2	0.544	0.050	0.00	1.28	91.0	0.545	0.064	0.00	94.9
450	0.400	0.408	0.059	0.01	93.5	0.402	0.050	0.00	1.38	87.5	0.403	0.067	0.00	94.4
							$\pi_r = 0.6$	$c = 30\%$						
100	0.809	0.811	0.041	0.00	93.1	0.810	0.039	0.00	1.09	91.3	0.801	0.045	0.01	93.2
300	0.541	0.544	0.059	0.00	93.6	0.541	0.052	0.00	1.28	89.9	0.521	0.069	0.02	90.1
450	0.400	0.407	0.062	0.01	93.5	0.401	0.053	0.00	1.38	87.0	0.375	0.074	0.03	85.3
							$\pi_r = 0.6$	$c = 50\%$						
100	0.809	0.813	0.042	0.00	94.0	0.811	0.040	0.00	1.08	90.0	0.764	0.046	0.04	76.8
300	0.541	0.551	0.063	0.01	91.9	0.544	0.056	0.00	1.27	91.4	0.430	0.070	0.11	57.5
450	0.400	0.409	0.070	0.01	92.3	0.399	0.060	0.00	1.36	86.4	0.248	0.070	0.15	45.9
							$\pi_r = 0.6$	$c = 60\%$						
100	0.809	0.813	0.042	0.00	92.3	0.811	0.041	0.00	1.08	90.5	0.717	0.046	0.09	53.0
300	0.541	0.550	0.067	0.01	92.6	0.543	0.059	0.00	1.26	87.4	0.319	0.066	0.22	29.2
450	0.400	0.412	0.079	0.01	90.4	0.397	0.070	0.00	1.29	84.6	0.108	0.039	0.29	18.0
							$\pi_r = 0.8$	$c = 0\%$						
100	0.886	0.887	0.034	0.00	91.3	0.885	0.032	0.00	1.19	93.8	0.886	0.036	0.00	91.9
300	0.657	0.659	0.057	0.00	93.1	0.655	0.047	0.00	1.45	87.9	0.656	0.061	0.00	93.9
450	0.505	0.509	0.062	0.00	93.6	0.504	0.050	0.00	1.54	86.6	0.505	0.067	0.00	94.4
							$\pi_r = 0.8$	$c = 10\%$						
100	0.886	0.887	0.035	0.00	92.4	0.886	0.032	0.00	1.20	89.1	0.886	0.036	0.00	93.2
300	0.657	0.660	0.058	0.00	94.8	0.655	0.048	0.00	1.45	86.7	0.656	0.063	0.00	94.7
450	0.505	0.510	0.063	0.01	94.4	0.503	0.051	0.00	1.53	86.1	0.504	0.070	0.00	95.2
							$\pi_r = 0.8$	$c = 30\%$						
100	0.886	0.887	0.035	0.00	90.5	0.886	0.032	0.00	1.20	88.5	0.883	0.037	0.00	90.8
300	0.657	0.658	0.060	0.00	94.2	0.654	0.049	0.00	1.45	86.3	0.645	0.067	0.01	92.9
450	0.505	0.511	0.066	0.01	93.7	0.504	0.053	0.00	1.54	85.5	0.492	0.076	0.01	91.4
							$\pi_r = 0.8$	$c = 50\%$						
100	0.886	0.889	0.035	0.00	89.6	0.887	0.032	0.00	1.20	88.6	0.868	0.038	0.01	85.7
300	0.657	0.661	0.062	0.00	94.9	0.657	0.051	0.00	1.46	87.2	0.597	0.070	0.06	77.3
450	0.505	0.512	0.071	0.01	93.6	0.505	0.057	0.00	1.53	83.6	0.419	0.080	0.08	71.4
							$\pi_r = 0.8$	$c = 60\%$						
100	0.886	0.887	0.035	0.00	91.4	0.886	0.032	0.00	1.20	88.7	0.844	0.039	0.04	73.7
300	0.657	0.660	0.064	0.00	93.2	0.655	0.053	0.00	1.45	85.8	0.529	0.071	0.12	55.2
450	0.505	0.511	0.075	0.01	93.0	0.502	0.061	0.00	1.51	85.3	0.323	0.076	0.18	41.0

TABLE 1.2: Simulation results for $n = 200$

		WRSE				**WKM**					**LDT**			
t	$S_1(t)$	$\hat{S}_1(t)$	SE	Bias	CP	$\hat{S}_1(t)$	SE	Bias	RE	CP	$\hat{S}_1(t)$	SE	Bias	CP
							$\pi_r = 0.2$	$c = 0\%$						
100	0.655	0.655	0.034	0.00	94.5	0.654	0.034	0.00	1.01	93.9	0.655	0.036	0.00	95.9
300	0.309	0.311	0.034	0.00	94.1	0.309	0.033	0.00	1.12	94.8	0.309	0.038	0.00	96.0
450	0.190	0.192	0.030	0.00	92.1	0.190	0.028	0.00	1.22	91.8	0.190	0.035	0.00	95.3
							$\pi_r = 0.2$	$c = 10\%$						
100	0.655	0.657	0.034	0.00	94.1	0.656	0.034	0.00	1.02	94.6	0.655	0.036	0.00	95.0
300	0.309	0.312	0.036	0.00	95.0	0.310	0.034	0.00	1.12	93.2	0.308	0.041	0.00	95.6
450	0.190	0.193	0.032	0.00	94.1	0.190	0.029	0.00	1.22	93.0	0.188	0.038	0.00	94.3
							$\pi_r = 0.2$	$c = 30\%$						
100	0.655	0.655	0.035	0.00	95.3	0.654	0.035	0.00	1.02	94.8	0.628	0.037	0.02	85.1
300	0.309	0.311	0.039	0.00	94.5	0.308	0.037	0.00	1.11	94.7	0.256	0.042	0.05	64.3
450	0.190	0.194	0.037	0.00	94.2	0.190	0.034	0.00	1.20	93.6	0.129	0.037	0.06	53.5
							$\pi_r = 0.2$	$c = 50\%$						
100	0.655	0.657	0.036	0.00	93.5	0.656	0.036	0.00	1.01	93.7	0.550	0.036	0.10	28.6
300	0.309	0.314	0.048	0.01	94.3	0.309	0.046	0.00	1.08	93.8	0.099	0.030	0.21	8.50
450	0.190	0.241	0.054	0.05	75.5	0.226	0.054	0.04	0.95	74.9	0.000	0.000	0.19	0.00
							$\pi_r = 0.2$	$c = 60\%$						
100	0.655	0.657	0.038	0.00	94.7	0.656	0.037	0.00	1.01	93.5	0.486	0.036	0.16	9.50
300	0.309	0.332	0.060	0.02	84.8	0.321	0.062	0.01	0.95	84.6	0.000	0.000	0.30	0.00
450	0.190	0.332	0.060	0.14	34.5	0.321	0.062	0.13	0.95	38.5	0.000	0.000	0.19	0.00
							$\pi_r = 0.4$	$c = 0\%$						
100	0.732	0.733	0.032	0.00	95.8	0.731	0.031	0.00	1.06	93.1	0.732	0.034	0.00	97.3
300	0.425	0.428	0.038	0.00	95.3	0.425	0.035	0.00	1.21	94.4	0.426	0.043	0.00	96.6
450	0.295	0.298	0.037	0.00	94.0	0.295	0.032	0.00	1.31	91.8	0.296	0.042	0.00	95.6
							$\pi_r = 0.4$	$c = 10\%$						
100	0.732	0.732	0.032	0.00	93.9	0.731	0.031	0.00	1.05	93.9	0.731	0.035	0.00	95.5
300	0.425	0.427	0.039	0.00	94.1	0.424	0.036	0.00	1.20	94.1	0.424	0.045	0.00	95.4
450	0.295	0.296	0.038	0.00	93.5	0.293	0.033	0.00	1.31	91.5	0.293	0.044	0.00	95.7
							$\pi_r = 0.4$	$c = 30\%$						
100	0.732	0.733	0.033	0.00	95.4	0.732	0.032	0.00	1.05	92.9	0.717	0.036	0.01	92.0
300	0.425	0.428	0.041	0.00	94.4	0.426	0.037	0.00	1.21	93.2	0.393	0.049	0.03	83.4
450	0.295	0.297	0.041	0.00	94.5	0.294	0.036	0.00	1.31	92.7	0.254	0.050	0.04	76.6
							$\pi_r = 0.4$	$c = 50\%$						
100	0.732	0.732	0.034	0.00	95.0	0.730	0.033	0.00	1.05	92.9	0.643	0.036	0.08	35.7
300	0.425	0.425	0.047	0.00	93.6	0.419	0.043	0.01	1.20	93.1	0.235	0.045	0.19	14.2
450	0.295	0.298	0.055	0.00	92.8	0.287	0.050	0.01	1.25	92.2	0.064	0.024	0.23	10.2
							$\pi_r = 0.4$	$c = 60\%$						
100	0.732	0.733	0.034	0.00	93.5	0.732	0.034	0.00	1.05	93.9	0.589	0.036	0.14	13.9
300	0.425	0.429	0.053	0.00	94.5	0.423	0.049	0.00	1.18	93.7	0.120	0.034	0.30	4.80
450	0.295	0.353	0.063	0.05	72.7	0.339	0.060	0.04	1.11	71.0	0.000	0.000	0.29	0.00
							$\pi_r = 0.6$	$c = 0\%$						
100	0.809	0.809	0.029	0.00	93.4	0.809	0.028	0.00	1.11	92.1	0.809	0.031	0.00	93.9
300	0.541	0.544	0.040	0.00	93.8	0.542	0.035	0.00	1.30	92.6	0.543	0.044	0.00	95.7
450	0.400	0.403	0.041	0.00	93.4	0.400	0.038	0.00	1.42	91.9	0.400	0.045	0.00	94.1
							$\pi_r = 0.6$	$c = 10\%$						
100	0.809	0.810	0.029	0.00	94.7	0.808	0.028	0.00	1.10	94.6	0.809	0.031	0.00	96.4
300	0.541	0.542	0.041	0.00	94.8	0.538	0.036	0.00	1.31	93.9	0.539	0.046	0.00	96.7
450	0.400	0.401	0.042	0.00	94.8	0.397	0.037	0.00	1.42	90.1	0.397	0.048	0.00	95.1
							$\pi_r = 0.6$	$c = 30\%$						
100	0.809	0.812	0.029	0.00	93.6	0.810	0.028	0.00	1.10	92.1	0.805	0.032	0.00	93.6
300	0.541	0.543	0.042	0.00	94.1	0.540	0.037	0.00	1.31	93.4	0.525	0.050	0.02	91.0
450	0.400	0.404	0.045	0.00	93.6	0.400	0.038	0.00	1.42	92.4	0.381	0.054	0.02	88.9
							$\pi_r = 0.6$	$c = 50\%$						
100	0.809	0.809	0.030	0.00	94.0	0.809	0.029	0.00	1.11	94.6	0.763	0.033	0.05	68.1
300	0.541	0.542	0.045	0.00	95.6	0.539	0.040	0.00	1.31	93.7	0.430	0.052	0.11	45.5
450	0.400	0.403	0.051	0.00	94.6	0.398	0.049	0.00	1.42	91.8	0.257	0.055	0.14	36.8
							$\pi_r = 0.6$	$c = 60\%$						
100	0.809	0.810	0.030	0.00	93.3	0.809	0.029	0.00	1.11	92.7	0.718	0.034	0.09	32.3
300	0.541	0.542	0.048	0.00	93.8	0.538	0.042	0.00	1.31	92.5	0.321	0.051	0.22	13.8
450	0.400	0.403	0.059	0.00	92.9	0.396	0.050	0.01	1.39	90.7	0.116	0.038	0.28	9.70
							$\pi_r = 0.8$	$c = 0\%$						
100	0.886	0.886	0.025	0.00	93.9	0.885	0.023	0.00	1.22	93.9	0.886	0.026	0.00	93.9
300	0.657	0.658	0.041	0.00	94.9	0.656	0.034	0.00	1.48	93.3	0.657	0.043	0.00	94.6
450	0.505	0.507	0.044	0.00	94.2	0.503	0.035	0.00	1.57	92.2	0.505	0.048	0.00	94.5
							$\pi_r = 0.8$	$c = 10\%$						
100	0.886	0.886	0.025	0.00	92.1	0.886	0.022	0.00	1.23	91.4	0.886	0.026	0.00	92.6
300	0.657	0.659	0.041	0.00	94.2	0.657	0.034	0.00	1.49	93.7	0.658	0.044	0.00	95.0
450	0.505	0.507	0.045	0.00	94.7	0.504	0.038	0.00	1.58	92.9	0.504	0.049	0.00	95.4
							$\pi_r = 0.8$	$c = 30\%$						
100	0.886	0.886	0.025	0.00	92.4	0.886	0.023	0.00	1.23	90.3	0.883	0.027	0.00	92.2
300	0.657	0.657	0.043	0.00	93.3	0.655	0.035	0.00	1.49	90.5	0.647	0.048	0.01	93.0
450	0.505	0.508	0.047	0.00	92.6	0.504	0.038	0.00	1.57	90.0	0.494	0.054	0.01	91.8
							$\pi_r = 0.8$	$c = 50\%$						
100	0.886	0.886	0.025	0.00	92.3	0.885	0.023	0.00	1.23	91.5	0.867	0.027	0.01	85.3
300	0.657	0.660	0.044	0.00	94.8	0.658	0.036	0.00	1.49	93.8	0.603	0.051	0.05	77.4
450	0.505	0.507	0.051	0.00	94.5	0.503	0.049	0.00	1.58	93.2	0.425	0.059	0.08	66.7
							$\pi_r = 0.8$	$c = 60\%$						
100	0.886	0.887	0.025	0.00	91.6	0.886	0.023	0.00	1.23	91.8	0.847	0.028	0.03	68.2
300	0.657	0.658	0.046	0.00	93.4	0.655	0.038	0.00	1.49	93.0	0.539	0.052	0.11	43.7
450	0.505	0.509	0.055	0.00	93.0	0.504	0.048	0.00	1.58	92.8	0.338	0.059	0.16	32.7

TABLE 1.3: Simulation results for $n = 400$

		WRSE				**WKM**				**LDT**			
t	$S_1(t)$	$\hat{S}_1(t)$	SE	Bias	CP	$\hat{S}_1(t)$	SE	Bias	RE	CP	$\hat{S}_1(t)$	SE	Bias	CP
							$\pi_r = 0.2$	$c = 0\%$						
100	0.655	0.655	0.024	0.00	95.3	0.655	0.024	0.00	1.03	94.5	0.655	0.025	0.00	96.2
300	0.309	0.310	0.024	0.00	95.3	0.309	0.023	0.00	1.12	92.9	0.309	0.027	0.00	97.4
450	0.190	0.191	0.022	0.00	96.2	0.190	0.020	0.00	1.22	92.5	0.189	0.024	0.00	97.1
							$\pi_r = 0.2$	$c = 10\%$						
100	0.655	0.656	0.024	0.00	95.2	0.655	0.024	0.00	1.02	94.5	0.655	0.026	0.00	96.0
300	0.309	0.311	0.025	0.00	94.2	0.310	0.024	0.00	1.12	93.5	0.309	0.029	0.00	95.2
450	0.190	0.192	0.023	0.00	94.9	0.190	0.021	0.00	1.22	93.4	0.189	0.027	0.00	94.7
							$\pi_r = 0.2$	$c = 30\%$						
100	0.655	0.656	0.025	0.00	95.5	0.656	0.024	0.00	1.02	95.8	0.632	0.026	0.02	80.8
300	0.309	0.310	0.028	0.00	94.1	0.309	0.026	0.00	1.13	94.4	0.262	0.031	0.05	55.9
450	0.190	0.192	0.027	0.00	93.4	0.190	0.024	0.00	1.22	93.1	0.135	0.029	0.06	46.1
							$\pi_r = 0.2$	$c = 40\%$						
100	0.655	0.656	0.026	0.00	96.1	0.655	0.026	0.00	1.02	95.2	0.554	0.026	0.10	10.8
300	0.309	0.310	0.035	0.00	94.4	0.307	0.033	0.00	1.12	95.0	0.105	0.025	0.20	3.60
450	0.190	0.231	0.042	0.04	73.8	0.223	0.042	0.03	1.03	72.8	0.000	0.000	0.19	0.00
							$\pi_r = 0.2$	$c = 60\%$						
100	0.655	0.656	0.027	0.00	93.9	0.654	0.026	0.00	1.02	94.3	0.489	0.026	0.17	3.30
300	0.309	0.327	0.047	0.02	85.7	0.319	0.048	0.01	1.00	86.0	0.000	0.000	0.30	0.00
450	0.190	0.327	0.047	0.13	22.2	0.319	0.048	0.13	1.00	26.2	0.000	0.000	0.19	0.00
							$\pi_r = 0.4$	$c = 0\%$						
100	0.732	0.731	0.023	0.00	94.6	0.731	0.022	0.00	1.06	94.3	0.731	0.024	0.00	95.9
300	0.425	0.425	0.027	0.00	94.3	0.424	0.025	0.00	1.21	93.3	0.424	0.030	0.00	96.2
450	0.295	0.296	0.026	0.00	95.1	0.294	0.023	0.00	1.32	92.7	0.295	0.029	0.00	96.5
							$\pi_r = 0.4$	$c = 10\%$						
100	0.732	0.732	0.023	0.00	93.0	0.731	0.022	0.00	1.06	94.8	0.731	0.025	0.00	94.5
300	0.425	0.427	0.028	0.00	94.9	0.426	0.025	0.00	1.21	93.2	0.425	0.032	0.00	96.4
450	0.295	0.297	0.027	0.00	95.7	0.295	0.023	0.00	1.32	92.1	0.295	0.032	0.00	95.9
							$\pi_r = 0.4$	$c = 30\%$						
100	0.732	0.734	0.023	0.00	93.9	0.733	0.023	0.00	1.06	93.1	0.720	0.026	0.01	89.4
300	0.425	0.426	0.029	0.00	94.7	0.424	0.027	0.00	1.22	93.0	0.396	0.036	0.03	80.8
450	0.295	0.296	0.030	0.00	95.5	0.294	0.026	0.00	1.33	92.9	0.259	0.037	0.04	73.8
							$\pi_r = 0.4$	$c = 40\%$						
100	0.732	0.733	0.024	0.00	95.7	0.732	0.023	0.00	1.06	94.8	0.647	0.026	0.08	21.2
300	0.425	0.427	0.033	0.00	95.7	0.425	0.030	0.00	1.22	93.5	0.241	0.034	0.18	6.05
450	0.295	0.300	0.040	0.01	92.5	0.296	0.035	0.00	1.33	92.0	0.072	0.024	0.22	4.80
							$\pi_r = 0.4$	$c = 60\%$						
100	0.732	0.733	0.024	0.00	95.4	0.732	0.024	0.00	1.06	93.9	0.595	0.026	0.13	5.30
300	0.425	0.427	0.038	0.00	94.0	0.425	0.035	0.00	1.22	90.5	0.132	0.029	0.29	2.70
450	0.295	0.341	0.051	0.04	73.8	0.332	0.047	0.04	1.87	72.1	0.000	0.000	0.29	0.00
							$\pi_r = 0.6$	$c = 0\%$						
100	0.809	0.810	0.021	0.00	94.8	0.810	0.020	0.00	1.11	94.5	0.810	0.022	0.00	95.6
300	0.541	0.541	0.029	0.00	95.5	0.540	0.025	0.00	1.32	93.2	0.540	0.031	0.00	96.6
450	0.400	0.402	0.029	0.00	96.4	0.400	0.024	0.00	1.43	92.4	0.400	0.032	0.00	96.8
							$\pi_r = 0.6$	$c = 10\%$						
100	0.809	0.810	0.021	0.00	95.3	0.809	0.020	0.00	1.11	94.9	0.809	0.022	0.00	96.3
300	0.541	0.542	0.029	0.00	94.7	0.541	0.025	0.00	1.32	93.7	0.541	0.032	0.00	95.7
450	0.400	0.402	0.030	0.00	93.7	0.400	0.025	0.00	1.43	92.0	0.400	0.034	0.00	94.7
							$\pi_r = 0.6$	$c = 30\%$						
100	0.809	0.810	0.021	0.00	94.3	0.810	0.020	0.00	1.12	93.0	0.803	0.023	0.01	94.1
300	0.541	0.542	0.030	0.00	94.0	0.541	0.026	0.00	1.32	93.0	0.526	0.036	0.02	90.9
450	0.400	0.402	0.032	0.00	94.9	0.401	0.027	0.00	1.44	93.0	0.381	0.039	0.02	88.4
							$\pi_r = 0.6$	$c = 50\%$						
100	0.809	0.810	0.021	0.00	95.2	0.809	0.020	0.00	1.11	94.5	0.766	0.024	0.04	54.3
300	0.541	0.542	0.032	0.00	95.1	0.541	0.028	0.00	1.33	93.9	0.438	0.038	0.10	31.0
450	0.400	0.402	0.036	0.00	95.1	0.400	0.030	0.00	1.46	93.8	0.265	0.041	0.13	23.0
							$\pi_r = 0.6$	$c = 60\%$						
100	0.809	0.809	0.022	0.00	95.2	0.809	0.021	0.00	1.12	93.9	0.719	0.024	0.09	14.0
300	0.541	0.543	0.035	0.00	93.4	0.541	0.030	0.00	1.34	93.1	0.327	0.038	0.21	5.00
450	0.400	0.403	0.043	0.00	94.7	0.399	0.040	0.00	1.46	92.6	0.122	0.034	0.27	3.00
							$\pi_r = 0.8$	$c = 0\%$						
100	0.886	0.887	0.018	0.00	92.9	0.886	0.016	0.00	1.24	92.6	0.887	0.018	0.00	93.2
300	0.657	0.657	0.029	0.00	94.3	0.656	0.024	0.00	1.50	93.9	0.657	0.031	0.00	95.2
450	0.505	0.506	0.032	0.00	95.8	0.505	0.028	0.00	1.59	91.7	0.505	0.034	0.00	96.2
							$\pi_r = 0.8$	$c = 10\%$						
100	0.886	0.886	0.018	0.00	94.1	0.886	0.016	0.00	1.25	93.5	0.886	0.018	0.00	95.1
300	0.657	0.658	0.029	0.00	94.5	0.657	0.024	0.00	1.50	94.8	0.657	0.031	0.00	95.7
450	0.505	0.507	0.032	0.00	94.7	0.505	0.028	0.00	1.59	92.9	0.505	0.035	0.00	94.8
							$\pi_r = 0.8$	$c = 30\%$						
100	0.886	0.886	0.018	0.00	94.3	0.886	0.016	0.00	1.24	94.8	0.884	0.019	0.00	93.7
300	0.657	0.659	0.030	0.00	95.4	0.657	0.025	0.00	1.51	94.4	0.652	0.034	0.01	94.5
450	0.505	0.507	0.034	0.00	95.4	0.506	0.029	0.00	1.60	93.3	0.497	0.039	0.01	93.2
							$\pi_r = 0.8$	$c = 50\%$						
100	0.886	0.887	0.018	0.00	94.1	0.886	0.016	0.00	1.24	93.9	0.869	0.020	0.01	83.2
300	0.657	0.659	0.032	0.00	95.8	0.658	0.026	0.00	1.51	94.3	0.605	0.037	0.05	65.8
450	0.505	0.509	0.037	0.00	95.0	0.507	0.034	0.00	1.61	93.8	0.431	0.043	0.07	57.1
							$\pi_r = 0.8$	$c = 60\%$						
100	0.886	0.887	0.018	0.00	94.2	0.886	0.016	0.00	1.25	93.9	0.851	0.020	0.04	55.8
300	0.657	0.659	0.033	0.00	94.7	0.658	0.027	0.00	1.52	91.6	0.550	0.038	0.11	27.9
450	0.505	0.509	0.039	0.00	96.0	0.507	0.031	0.00	1.62	90.7	0.351	0.045	0.15	18.9

1.7 Conclusion

In this chapter, adaptive treatment interventions are introduced and a brief review of counterfactuals is done. The model framework used in inference for dynamic treatment regimes is also introduced. We review three non-parametric estimators for survival distributions. All these estimators use the concept of inverse probability weighting. Patients who could have been randomized to a treatment policy of interest but end up in another treatment policy are regarded as missing in the treatment of interest. To address such missingness which happens because of the nature the trial is done, inverse weights are used. The first estimator, which we referred to as the LDT, uses two forms two forms of inverse probability weighting. The WRSE is an extension of the Nelson-Aalen estimator and the the WKM estimator is an extension of the Kaplan-Meier estimator. No comparative study has been done for the three methods. We provide a simulation study to compare the three methods in cases of extreme response rates and censoring rates. Three sample sizes are considered. The estimators are affected by response rates and censoring rates. The LDT estimator is drastically affected by low response rates and censoring rates. The other two estimators are affected by low response rates and high censoring rates but the impact is minimal. The WRSE and the WKPM estimator perform better and the WRSE has better coverage probabilities than the WKM estimator.

Chapter 2

The data

2.1 CALGB 19808 Study

In the Cancer and Leukemia Group B 19808 (CALGB 19808) study, 302 patients were randomized to receive induction chemotherapy regimens consisting of cytosine arabinoside (Ara-C;A), daunorubicin (D), and etoposide (E) without (ADE) or with (ADEP) PSC-833 (P) (Kolitz et al., 2010). The study was done to patients under the age of 60 with newly diagnosed acute myeloid leukemia. To be eligible, the patients should not have been previously treated for leukemia and be under the age of 60. The study was designed to compare the two induction chemotherapy regimens, ADE and ADEP, with both treatments given at their highest clinically feasible doses.

For the first stage, the main objective of the trial was to determine whether the use of the Pgp-modulating agent PSC-833 in the ADEP regimen improved overall survival and disease free survival compared to ADE only. The randomization between ADE and ADEP was done at 1:1 ratio. The analysis of the first stage data is reported in Kolitz et al. (2010). In both treatment arms, 75% of the patients achieved complete remission (CR). Complete remission was defined using the National Cancer Institute Workshop criteria (Cheson et al., 1990). The results of the first stage analysis showed no significant difference for the two treatments in terms of overall and disease free survival.

The 75% who achieved complete remission were further randomized to the second stage treatments namely recombinant interleukin-2 (rIL-2) and no rIL-2 (observation). The goal of the second stage was to assess the effect of rIL-2 immunotherapy on disease free survival. Neither disease free or overall survival was found to be significantly improved from the analysis of the second stage data. Patients who were alive and still in compete remission at the time of the last follow-up were considered as right-censored

observations. Overall survival was measured from the date of randomization to the treatments to death, patients still alive at the last follow-up were censored. The trial suffered setbacks in the second stage due to unexpected refusals by patients or their medical doctors to comply with protocol as preplanned. Figure 2.1, a CONSORT diagram taken from Kolitz et al. (2010), summarizes the study before the refusals.

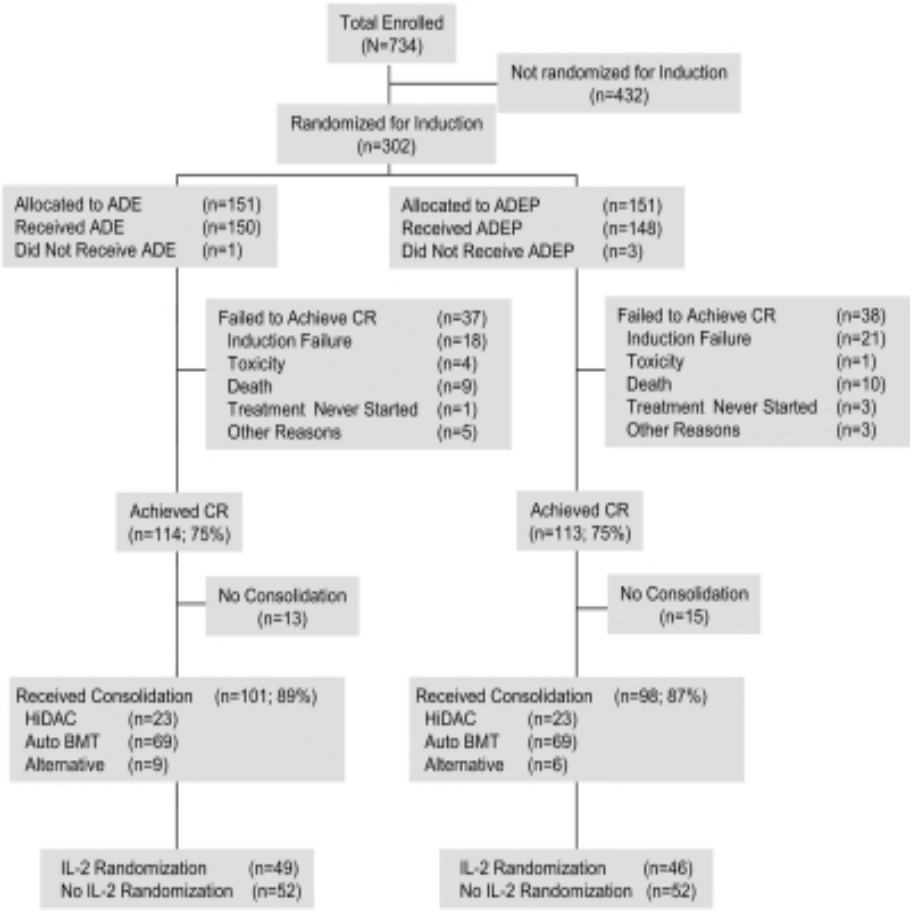


FIGURE 2.1: CONSORT diagram for the CALGB 19808 study.

2.2 Exploratory analysis

Since some of the patients refused to take the second stage treatments, we regard this as withdrawn consent and are regarded as non-responders. A treatment regime, say A_jB_k, means treating with A_j followed by B_k if the patient is eligible and consents to next stage therapy. Not only were the refusals the reasons patients went off treatment. We shall in turn summarize some of the other reasons below. The following table summarizes the numbers in both stages of the trial, that is, the actual numbers who eventually completed the second stage as per the protocol.

Table 2.1: Actual numbers

	First stage		Second stage	
	N	Response	Obs	rIL-2
ADE	151	114 (75%)	26	23
ADEP	151	113 (75%)	21	25

The second stage suffered heavily because of the refusals and the other reasons for patients to leave the treatment. The response rate also declined as some of these patients were then regarded as non responders in the analysis. Among those who had responded to the first stage treatments, 56 refused the second stage treatments and follow-up was done to them. Three patients refused with no followup done to them. Twenty three patients left for alternative therapy. Other patients were put off treatment for other reasons.

The first component of the survival mixture in (3.23) can come from any survival distribution deemed suitable, while the second component requires a parametric convolution of survival distributions. Since we can assume any distribution for the first component in the mixture, it is useful to check which survival distribution fits well the data for the non-responders in the two treatment arms. In Figure 2.2, we compared the fits from the assumed parametric distributions to the Kaplan-Meier estimator, where we used the exponential, Weibull and Gompertz distributions.

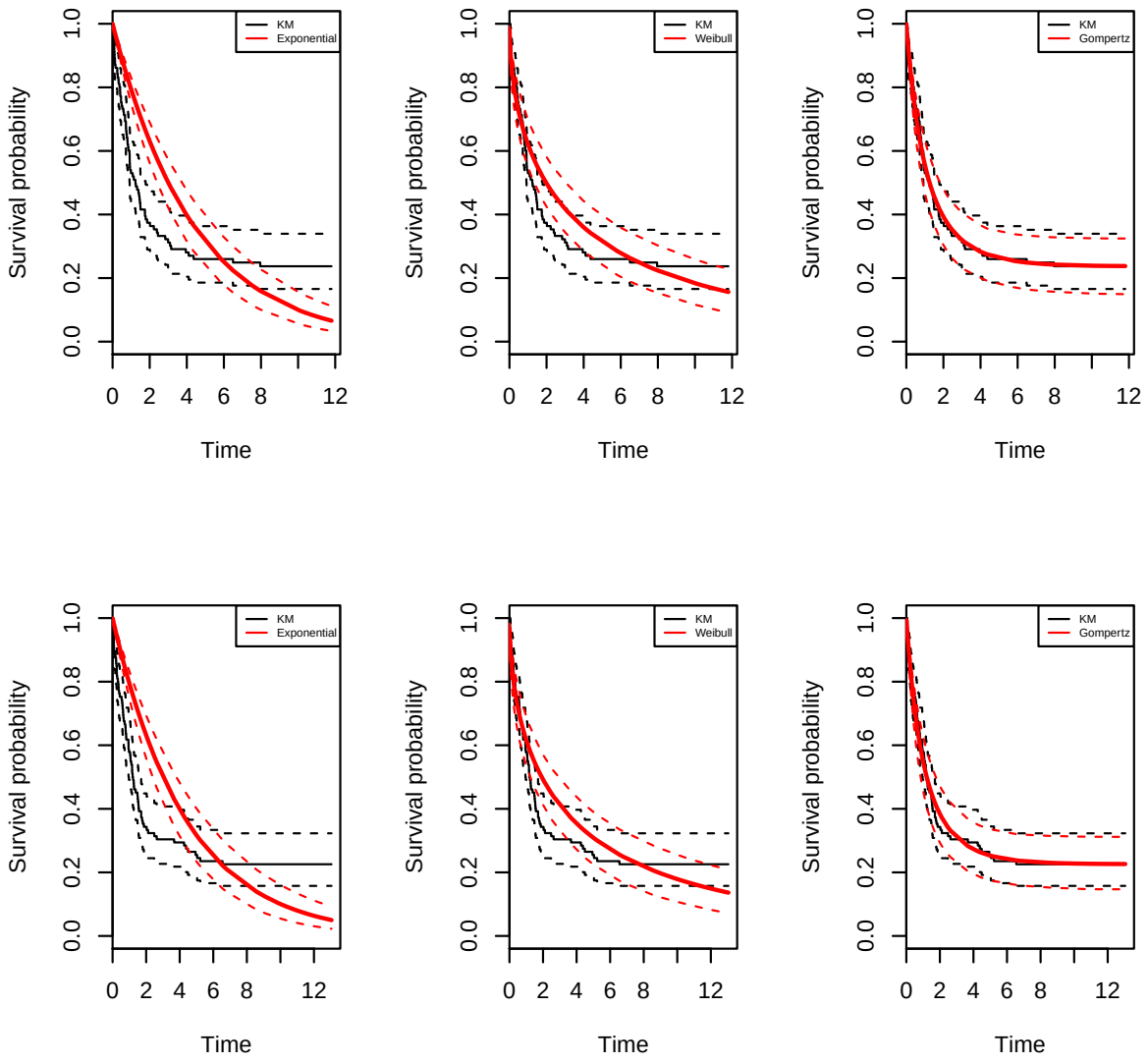


FIGURE 2.2: Fitting exponential, Weibull and Gompertz distributions to the non-responders data: Non-responders to ADE (first three graphs) and to ADEP (last three graphs).

The first three set of plots are from data on non-responders to ADE and the last three set of plots are from data from non-responders to ADEP. In both instances the exponential model fits poorly the data as there is big discrepancy between the curve and the one from the Kaplan-Meier estimator. The Weibull provides a better fit than the exponential but there is still a big difference from the graph of the Kaplan-Meier estimator. In both cases, the Gompertz distribution provides the best fit. Similar graphs for the times to response and the times from response to an event (T_i^R, T_{ki}^*) are

also shown in Figures 2.3 and 2.4 where we fitted survival curves for ADE-rIL-2 and ADEP-rIL-2. The graphs from the other treatment policies are similar.

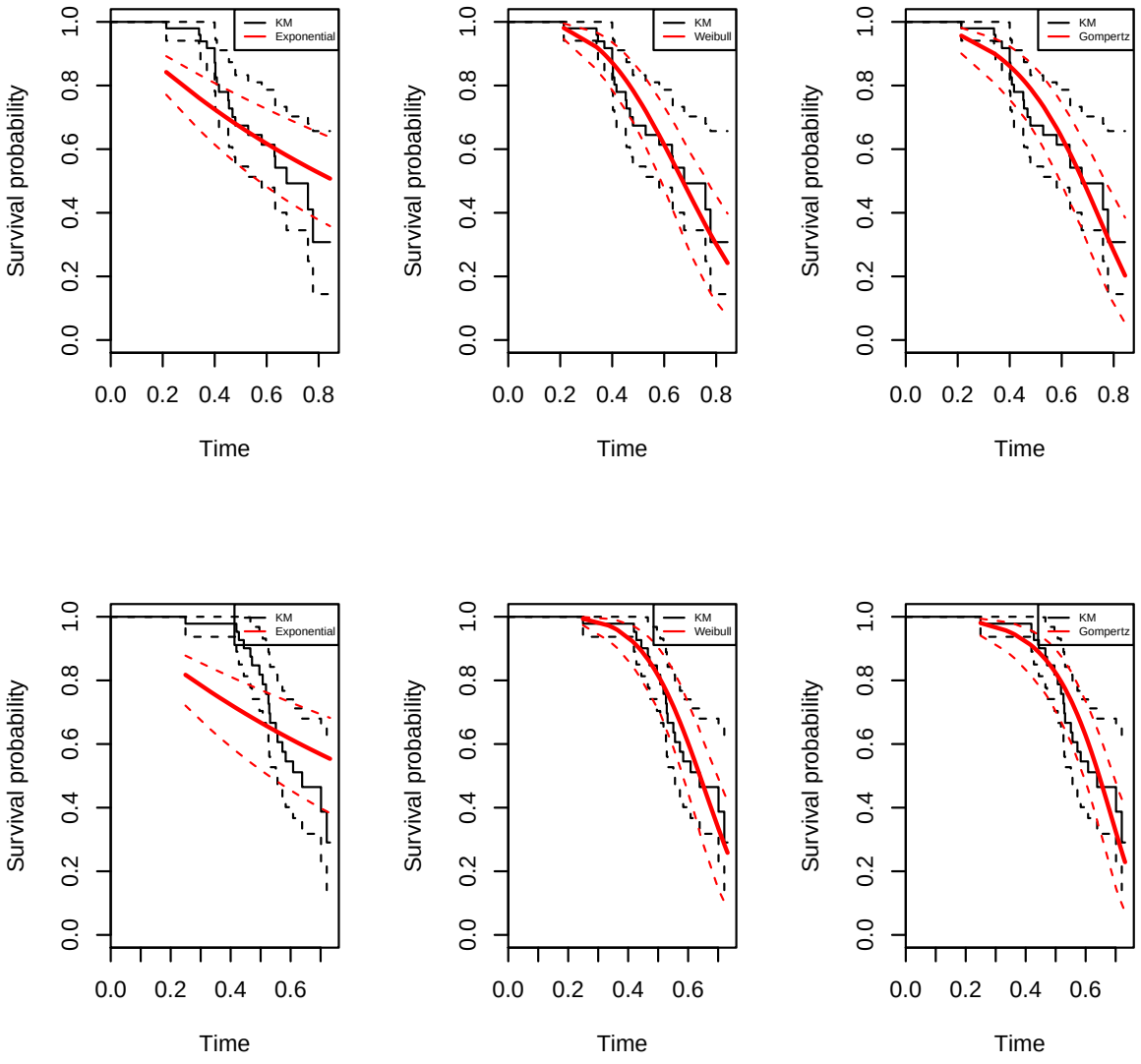


FIGURE 2.3: Comparison of parametric models: exponential, Weibull and Gompertz for times to response to ADE (on the top) and to ADEP (on the bottom)

.

Figure 2.3 shows the survival curves for the times to response to ADE and ADEP (T_i^R). In both cases, the exponential model provides a very poor fit. The Weibull model and the Gompertz model fits are reasonable in both cases though the Gompertz model fit is slightly better. Figure 2.4 shows the survival curves for rIL-2 under ADE and ADEP (T_1^*, T_2^*). Again the exponential model provides a poor fit when compared to the Kaplan-Meier curve. The Weibull model provides a better fit. The Gompertz

model provides the best fit in both instances. Survival curves for observation (OBS), also referred to as no rIL-2, under ADE and ADEP are similar and are not shown here.

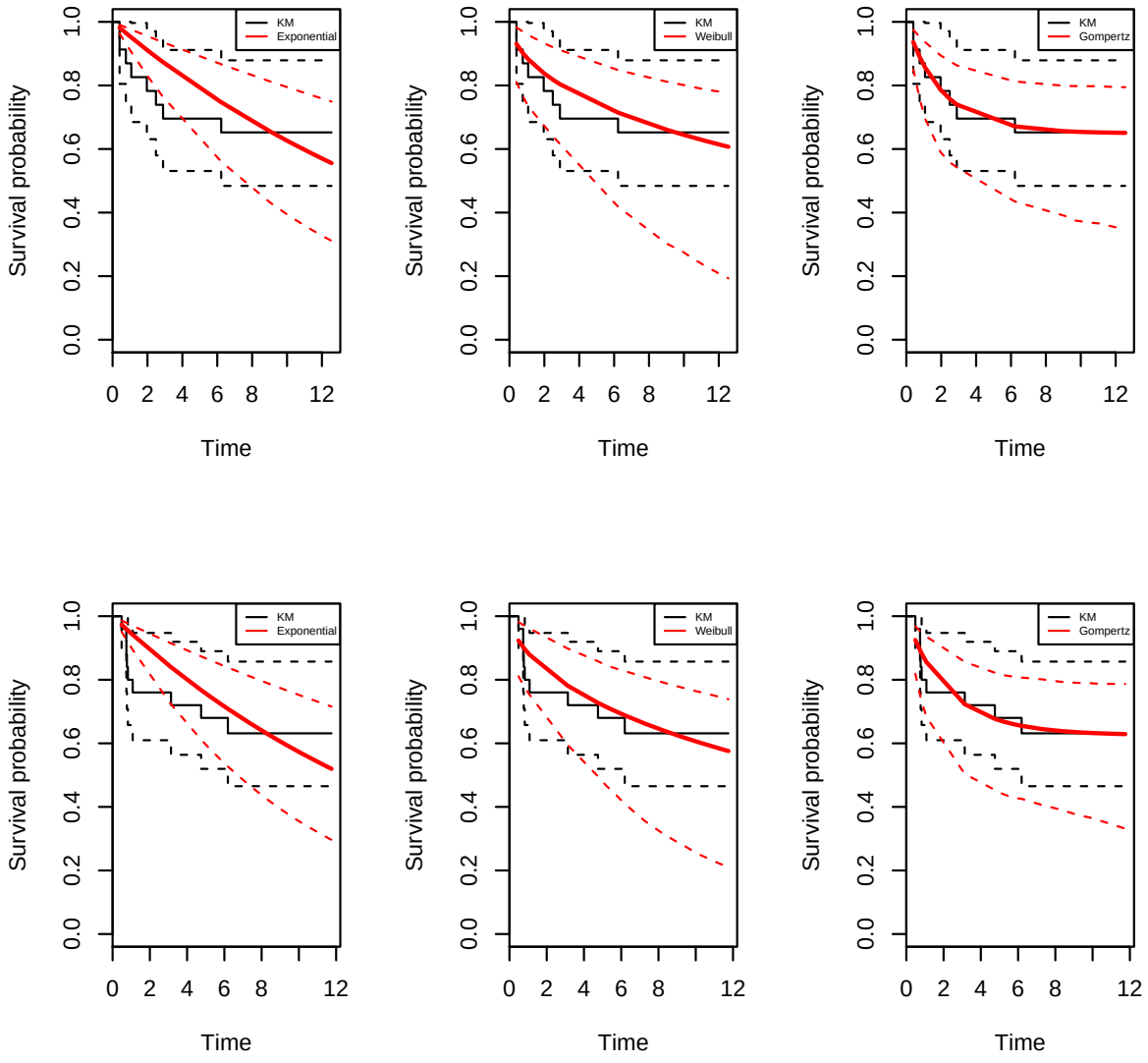


FIGURE 2.4: Comparison of parametric models: exponential, Weibull and Gompertz for rIL-2 under ADE (top) and ADEP (bottom).

2.2.1 Weighted analysis

In Figure 2.5, we show the survival curves for the four treatment policies embedded in the CALGB 19808 study. The curves were estimated using the WRSE. The four curves are clustered together and there seems to be no difference in the survival experiences of the patients in these treatment policies. Also the curves cross at different points, this

poses a challenge in the comparison of the survival curves. We take this problem further in Chapter 4.

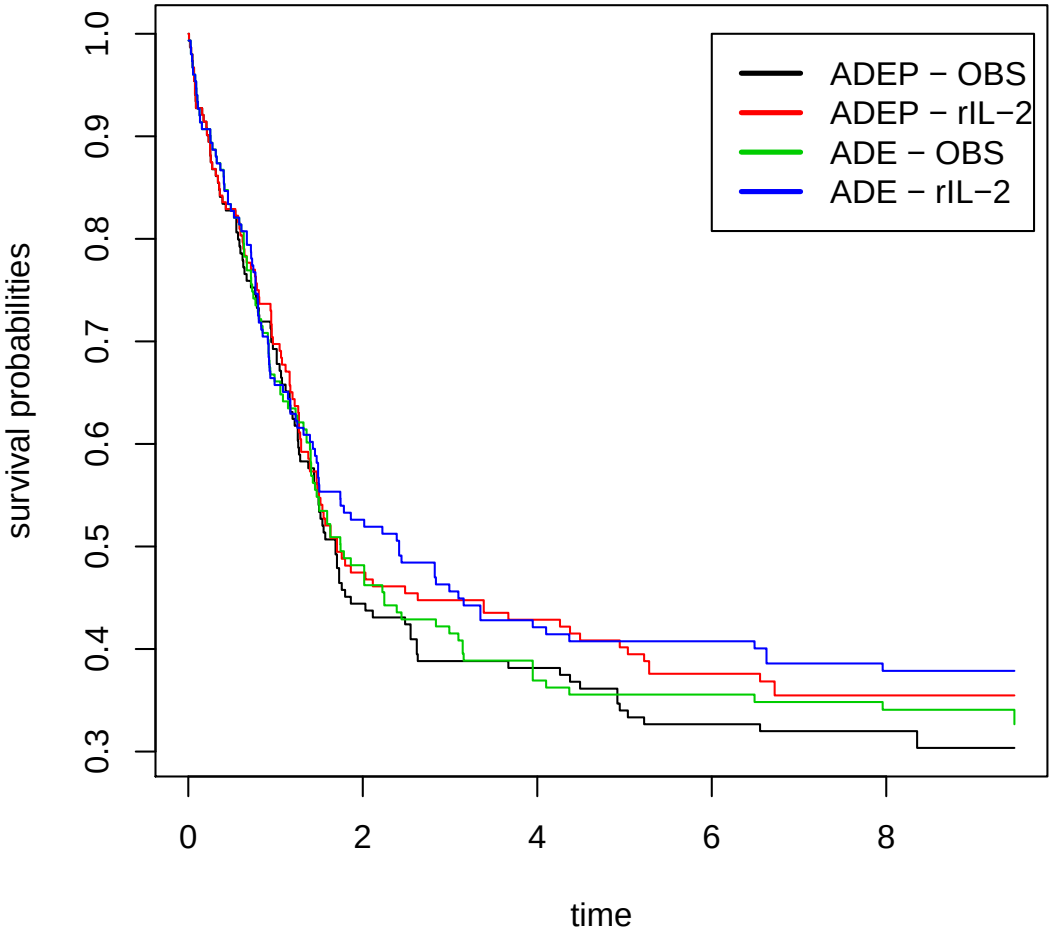


FIGURE 2.5: Survival curves for CALGB 19808.

Chapter 3

Estimating survival distributions for time-varying SMART designs

In addition to the nonparametric approaches for survival distributions in two-stage designs, some parametric approaches have been proposed in literature. Thall et al. (2002) developed a parametric approach for selecting the best strategy on the basis of the mean overall failure times. A Bayesian framework was proposed but the method suffers from overparametrization as so many parameters need to be estimated. Since the study by Thall et al. (2002) is Bayesian in nature we will not consider it further.

3.1 Parametric mixture approach

Wahed (2010) developed a likelihood based method for estimating the survival means for adaptive treatment strategies upon which inferences are made to compare different treatment policies. The development of this approach is also based on counterfactuals. Hereafter, we describe Wahed (2010) approach in a design where we consider first stage treatments A_j, $j = 1, 2$ and second stage treatments B_k, $k = 1, 2$. Let

$$T_{jki} = (1 - R_{ji})T_{j0i} + R_{ji}T^{**}_{jki},\ j = 1, 2; k = 1, 2, \tag{3.1}$$

and the observed survival time is

$$T_i = \sum_{i=1}^{n} X_{ji} \left\{(1 - R_{ji})T_{j0i} + R_{ji}Z_{ki}T^{**}_{jki}\right\},\ j = 1, 2; k = 1, 2,$$

where X_{ji} is the first treatment indicator and T^{**}_{jki} is the overall survival time for the ith patient assigned to treatment policy A_jB_k. Note that this survival time is different

from T^*_{jk} previously defined in Chapter 2. T^{**}_{jki} is the total survival time from first randomization to an event while T^*_{jk} is the time from second randomization to an event. Define $X_2 = 1 - X_1$ and $Z_2 = 1 - Z_1$. It is further assumed that, by design, the randomization probabilities are independent of the observed data.

To construct the likelihood for the observed data, probability models are assumed for the counterfactual times. Let $E[h(T^{**}_{jki})] \equiv \gamma_{jk},\ j = 1, 2; k = 0, 1, 2.$, where $h(.)$ is some function based on the data. Noting that the survival times for the treatment policy A_jB_k is a mixture of two survival counterfactual variables, the expected value for the treatment policy A_jB_k can be written as

$$\mu_{jk} = (1 - \pi_{rj})\gamma_{j0} + \pi_{rj}\gamma_{jk}, \quad j = 1, 2; k = 1, 2, \tag{3.2}$$

where π_{rj} is the proportion of responders in arm A_j, $j = 1, 2$. Let $X_{ji} \sim \text{Bernoulli}(\pi_{xj})$, $R_{ji}|X_{ji} \sim \text{Bernoulli}(\pi_{rj})$, $Z_{ki}|R_i \sim \text{Bernoulli}(\pi_{zk})$, $T^{**}_{jk} \sim f(.; \theta_{jk})\ j = 1, 2; k = 1, 2$ and $T_{j0} \sim f(.; \theta_{j0})$, $j = 1, 2$. π_{xj} is the proportion of subjects assigned to A_j, $j = 1, 2$ and π_{zk} is the proportion of subjects assigned to B_k, $k = 1, 2$. We define $\pi_{z2} = 1 - \pi_{z1}$ and $\pi_{x2} = 1 - \pi_{x1}$. Let r_i be a realization of R_i and δ be a realization of Δ, with right censoring, the observed data are $D_i = (X_{1i}, R_iZ_{1i}, U_i, \Delta_i)$ and the full likelihood is

$$L(\theta, \pi; \{D_i\}_{i=1}^n) = L^1(\pi; \{x_{1i}, r_i, r_iz_{1i}\}_{i=1}^n)L^2(\theta; \{x_{1i}, r_i, r_iz_{1i}, u_i, \delta_i\}_{i=1}^n, \tag{3.3}$$

where $\pi = (\pi_{r1}, \pi_{r2}, \pi_{x1}, \pi_{z1})$, $\theta = (\theta_{jk}\ j = 1, 2; k = 0, 1, 2)$,

$$L^1(\pi; \{x_{1i}, r_i, r_iz_{1i}\}_{i=1}^n) = \prod_{i=1}^{n} b(x_{1i}; \pi_{x1}) \prod_{j=1}^{2} \{b(r_i; \pi_{rj})b(z_{1i}; \pi_{z1})\}^{x_{ji}}, \tag{3.4}$$

where $b(.; p)$ is the probability mass function for a Bernoulli random variable with success probability p and

$$\begin{aligned} L^2(\theta; \{x_{1i}, r_i, r_iz_{1i}, u_i, \delta_i\}_{i=1}^n = \prod_{i=1}^{n} \prod_{j=1}^{2} \Bigg(\Bigg[\prod_{k=1}^{2} \{f_{jk}(u_i; \theta_{jk})^{\delta_i} S_{jk}(u_i; \theta_{jk})^{1-\delta_i}\}^{z_{ki}} \Bigg]^{r_i} \\ \times \{f_{jk}(u_i; \theta_{j0})^{\delta_i} S_{j0}(u_i; \theta_{j0})^{1-\delta_i}\}^{1-r_i} \Bigg)^{x_j i}. \end{aligned} \tag{3.5}$$

The likelihood factorizes into two components: the likelihood contribution for π and the likelihood contribution for θ. To estimate survival distributions for the treatment

strategies one replaces the means in (3.2) with survival functions to get

$$\mathcal{S}_{jk}(u) = (1 - \pi_{rj})S_{j0}(u) + \pi_{rj}S_{jk}(u),\ j = 1, 2; k = 1, 2. \tag{3.6}$$

This is a well known result from the theory of mixture distributions (McLachlan and McGiffin, 1993).

3.2 Parametric approach for time-varying SMART designs

We introduced time-varying SMART designs in Chapter 1. For simplicity, let us consider the treatment A_1 since results are similar for treatment A_2. Wahed (2010), defined using counterfactuals, the survival time for patient i, if assigned to A_1B_k, as

$$T_{ki} = (1 - R_i)T_{0i} + R_iT_{ki}^{**}, \quad k = 1, 2. \tag{3.7}$$

This way of definition is not appropriate for two-stage time-varying SMART designs. In addition to T_{ki}^{**}, we need to consider another variable for responders, T_i^R, which is the time to response for the ith patient in the first stage. An example of a time-varying SMART design is shown in Figure 1.2. Since time to the first-stage response varies among the responders, it must be accounted for in the likelihood. In a time-varying SMART design the survival time should be defined as

$$T_{ki} = (1 - R_i)T_{0i} + R_i(T_i^R + T_{ki}^{*}), \quad k = 1, 2. \tag{3.8}$$

The observed survival time in this case is the sum of two random variables for the responders. One cannot put a single distribution on a sum as that could be theoretically incorrect. To solve this problem, we propose a parametric approach for the estimation of survival functions of treatment policies A_jB_k in the presence of a time-varying SMART design. This work follows the lines of Wahed (2010), extending some of the theory therein, to a more general setting.

3.2.1 Density of T_k

Let $\tilde{T}_{ki} = T_i^R + T_{ki}^*$, then $T_{ki} = (1 - R_i)T_{0i} + R_i\tilde{T}_{ki}$ for $k = 1, 2$. Let r be a realization of R, $r \in (0, 1)$. Then we can write

$$\begin{aligned} F_{T_k} &= P(T_k \leq t) = P([(1-r)T_0 + r\tilde{T}_k] \leq t) \\ &= \sum_{r\in(0,1)} P([(1-r)T_0 + r\tilde{T}_k] \leq t|R = r)P(R = r) \\ &= P(T_0 \leq t)P(R = 0) + P(\tilde{T}_k \leq t)P(R = 1) \\ &= (1 - \pi_r)P(T_0 \leq t) + \pi_r P(\tilde{T}_k \leq t), \quad k = 1, 2; \end{aligned}$$

where $P(R = 1) = \pi_r$. This leads to

$$f_{T_k}(t) = (1 - \pi_r)f_0(t) + \pi_r f_k(t), \quad k = 1, 2; \tag{3.9}$$

where $f_0(t)$ and $f_k(t)$ are the density functions of T_0 and $\tilde{T}_k$, respectively.

We note that $f_{\tilde{T}_k}(t)$ is obtained from a convolution of T^R and T_k^*. Using the relationship between a mixture density and the survival function (McLachlan and McGiffin, 1993), then the survivor function for treatment policy A_1B_K is given as;

$$S_{T_k}(t) = (1 - \pi_r)S_0(t) + \pi S_k(t), \tag{3.10}$$

where $S_0(t)$ and $S_k(t)$ are the survival functions of T_0 and $\tilde{T}_k$ respectively.

Example: Exponential model

Suppose that

$$\begin{aligned} T^R &\sim \lambda_R \exp(-\lambda_R t), \lambda_R > 0 \\ T_k^* &\sim \lambda_k \exp(-\lambda_k t), \lambda_k > 0, k = 1, 2. \end{aligned}$$

We are interested in the density of $\tilde{T}_k$:

$$\begin{aligned} f_{\tilde{T}_k}(\tilde{z}) &= \int_0^{\tilde{z}} \lambda_R \lambda_k e^{-\lambda_R t} e^{-\lambda_k(\tilde{z}-t)} dt \\ &= \int_0^{\tilde{z}} \lambda_R \lambda_k e^{-\lambda_R t} e^{-\lambda_k \tilde{z} + \lambda_k t} dt \\ &= \lambda_R \lambda_k e^{-\lambda_k \tilde{z}} \int_0^{\tilde{z}} e^{-(\lambda_R - \lambda_k)t} dt \\ &= \frac{\lambda_R \lambda_k}{\lambda_k - \lambda_R} e^{-\lambda_R \tilde{z}} + \frac{\lambda_R \lambda_k}{\lambda_R - \lambda_k} e^{-\lambda_k \tilde{z}}; \lambda_R, \lambda_k > 0, z \geq 0. \end{aligned} \tag{3.11}$$

Likewise, we can obtain the distribution function

$$\begin{aligned} F_{\tilde{T}_k}(\tilde{z}) &= P(\tilde{T}_k \leq \tilde{z}) \\ &= \frac{\lambda_R \lambda_k}{\lambda_k - \lambda_R} \int_0^{\tilde{z}} \left(e^{-\lambda_R t} - e^{-\lambda_k t} \right) dt \\ &= 1 + \frac{\lambda_R}{\lambda_k - \lambda_R} e^{-\lambda_k \tilde{z}} - \frac{\lambda_k}{\lambda_k - \lambda_R} e^{-\lambda_R \tilde{z}}. \end{aligned}$$

Consequently, the survival function for $\tilde{T}_k$ is

$$\begin{aligned} S_{\tilde{T}_k}(\tilde{z}) &= 1 - F_{\tilde{T}_k}(\tilde{z}) \\ &= 1 - \left[1 + \frac{\lambda_R}{\lambda_k - \lambda_R} e^{-\lambda_k \tilde{z}} - \frac{\lambda_k}{\lambda_k - \lambda_R} e^{-\lambda_R \tilde{z}} \right] \\ &= \frac{\lambda_k}{\lambda_k - \lambda_R} e^{-\lambda_R \tilde{z}} + \frac{\lambda_R}{\lambda_R - \lambda_k} e^{-\lambda_k \tilde{z}}. \end{aligned} \tag{3.12}$$

3.2.2 Likelihood and survival function

Supposed that the time-to-event is subject to right censoring. We assume that everyone's response status is always observed. To estimate the parameters needed for the survival distribution, we construct the likelihood for the observed data in a two-stage design.

The joint distribution of the data can be obtained as

$$
\begin{aligned}
&f(U_i = u_i, \Delta_i = \delta_i, R_i Z_i = r_i z_i | R_i = r_i) P(R_i = r_i) \\
&= f(U_i = u_i, \Delta_i = \delta_i | R_i Z_i = r_i z_i, R_i = r_i) P(R_i Z_i = r_i z_i | R_i = r_i) P(R_i = r_i) \\
&= \begin{cases}
f(U_{0i} = u_i, \Delta_i = \delta_i | R_i Z_i = 0, R_i = 0) P(R_i Z_i = 0 | R_i = 0) P(R_i = 0), & R_i = 0 \\
f(U_{1i} = u_i, \Delta_i = \delta_i | R_i Z_i = 1, R_i = 1) P(R_i Z_i = 1 | R_i = 1) P(R_i = 1), & R_i = 1, Z_i = 1 \\
f(U_{2i} = u_i, \Delta_i = \delta_i | R_i Z_i = 0, R_i = 1) P(R_i Z_i = 0 | R_i = 1) P(R_i = 1), & R_i = 1, Z_i = 0
\end{cases} \\
&= \begin{cases}
(1 - \pi_r) f_0(u_i)^{\delta_i} S_0(u_i)^{1-\delta_i} \\
\pi_r \pi_z f_1(u_i)^{\delta_i} S_1(u_i)^{1-\delta_i} \\
\pi_r (1 - \pi_z) f_2(u_i)^{\delta_i} S_2(u_i)^{1-\delta_i},
\end{cases}
\end{aligned}
\tag{3.13}
$$

where $P(Z_i = 1 | R_i = 1) = \pi_z$ which is the probability of being randomized to B_1 in the second stage. Clearly, $P(Z_i = 0 | R_i = 1) = 1 - \pi_z$ is the probability to be randomized to B_2.

Let O_i denote the observed data, $(r_i, r_i z_i, u_i, \delta_i)$ for patient i. Then, the full likelihood is

$$
\begin{aligned}
L(\theta, \pi; O) = \prod_{i=1}^{n} &[(1 - \pi_r) f_0(u_i)^{\delta_i} S_0(u_i)^{1-\delta_i}]^{1-r_i} \\
&\times \{[\pi_r \pi_z f_1(u_i)^{\delta_i} S_1(u_i)^{1-\delta_i}]^{z_i} . [\pi_r (1 - \pi_z) f_2(u_i)^{\delta_i} S_2(u_i)^{1-\delta_i}]^{1-z_i}\}^{r_i},
\end{aligned}
\tag{3.14}
$$

where $O = (O_1, O_2, ..., O_n)$, $\pi = (\pi_r, \pi_z)$ and $\theta = (\theta_R, \theta_k), k = 1, 2$. The likelihood factorizes into two parts, with one part depending only on the parameters π and the other part on the parameters θ;

$$
\begin{aligned}
L_1(\pi; O) &= (1 - \pi_r)^{\sum_{i=1}^{n} (1 - r_i)} . \pi_r^{\sum_{i=1}^{n} z_i r_i} . \pi_z^{\sum_{i=1}^{n} z_i r_i} . (1 - \pi_z)^{\sum_{i=1}^{n} r_i (1 - z_i)} . \pi_r^{\sum_{i=1}^{n} r_i (1 - z_i)} \\
&= (1 - \pi_r)^{\sum_{i=1}^{n} (1 - r_i)} . \pi_r^{\sum_{i=1}^{n} z_i r_i + \sum_{i=1}^{n} r_i - \sum_{i=1}^{n} z_i r_i} . \pi_z^{\sum_{i=1}^{n} z_i r_i} . (1 - \pi_z)^{\sum_{i=1}^{n} r_i (1 - z_i)} \\
&= (1 - \pi_r)^{\sum_{i=1}^{n} (1 - r_i)} . \pi_r^{\sum_{i=1}^{n} r_i} . \pi_z^{\sum_{i=1}^{n} z_i r_i} . (1 - \pi_z)^{\sum_{i=1}^{n} r_i (1 - z_i)} .
\end{aligned}
\tag{3.15}
$$

The corresponding log-likelihood is

$$l_1(\pi; O) = \log L_1(\pi; O) \tag{3.16}$$

$$= \sum_{i=1}^{n}(1 - r_i)\log(1 - \pi_r) + \sum_{i=1}^{n} r_i \log \pi_r + \sum_{i=1}^{n} z_i r_i \log \pi_z + \sum_{i=1}^{n} r_i(1 - z_i)\log(1 - \pi_z), \tag{3.17}$$

and

$$\begin{aligned} \frac{\partial l_1(\pi; O)}{\partial \pi_r} &= \frac{-\sum_{i=1}^{n}(1 - r_i)}{1 - \pi_r} + \frac{\sum_{i=1}^{n} r_i}{\pi_r} \\ \frac{\partial l_1(\pi; O)}{\partial \pi_z} &= \frac{\sum_{i=1}^{n} z_i r_i}{\pi_z} - \frac{\sum_{i=1}^{n} r_i(1 - z_i)}{1 - \pi_z}. \end{aligned} \tag{3.18}$$

Setting the two score equations from (3.18) to zero we get

$$\hat{\pi}_r = \frac{\sum_{i=1}^{n} r_i}{n} \tag{3.19}$$

$$\hat{\pi}_z = \frac{\sum_{i=1}^{n} r_i r_i}{\sum_{i=1}^{n} r_i}, \tag{3.20}$$

which are maximum likelihood estimators (MLEs) from $L_1(\pi; O)$. The likelihood for θ is,

$$L_2(\theta; O) = \prod_{i=1}^{n} [f_0(u_i)^{\delta_i} S_0(u_i)^{1-\delta_i}]^{1-r_i} \{[f_1(u_i)^{\delta_i} S_1(u_i)^{1-\delta_i}]^{z_i} . [f_2(u_i)^{\delta_i} S_2(u_i)^{1-\delta_i}]^{1-z_i}\}^{r_i}, \tag{3.21}$$

and the log-likelihood, $l_2(\theta; O) = \log L_2(\theta; O)$, becomes

$$\begin{aligned} l_2(\theta; O) = \sum_{i=1}^{n} \{&(1 - r_i)\log f_0(u_i)^{\delta_i} S_0(u_i)^{1-\delta_i} \\ &+ r_i z_i \log f_1(u_i)^{\delta_i} S_1(u_i)^{1-\delta_i} + r_i(1 - z_i)\log f_2(u_i)^{\delta_i} S_2(u_i)^{1-\delta_i}\}. \end{aligned} \tag{3.22}$$

To estimate the survival distributions for the treatment policy $A_1 B_k$, we propose using

$$\hat{S}_{A_1 B_k}(u) = (1 - \hat{\pi}_r)\hat{S}_0(u) + \hat{\pi}_r \hat{S}_k(u); \quad k = 1, 2, \tag{3.23}$$

where $\hat{S}_0(u)$ and $\hat{S}_k(u)$ are obtained by replacing the MLEs of θ in the parametric survival functions of $S_0(u)$ and $S_k(u)$. Estimating survival distributions for treatment policy $A_2 B_k$ follows analogously.

Example: Exponential model

Assuming exponential distribution we have

$$\begin{aligned} f_0(u) &= \lambda_0 e^{-\lambda_0 u} \\ f_1(u) &= \frac{\lambda_R \lambda_1}{\lambda_1 - \lambda_R} e^{-\lambda_R u} + \frac{\lambda_R \lambda_1}{\lambda_R - \lambda_1} e^{-\lambda_1 u} \\ f_2(u) &= \frac{\lambda_R \lambda_2}{\lambda_2 - \lambda_R} e^{-\lambda_R u} + \frac{\lambda_R \lambda_2}{\lambda_R - \lambda_2} e^{-\lambda_2 u}, \end{aligned}$$

and the log-likelihood becomes

$$\begin{aligned} l(\theta; O_i) = \sum_{i=1}^{n} \{ & (1 - r_i) \log[\lambda_0 e^{-\lambda_0 u_i}]^{\delta_i} [e^{-\lambda_0 u_i}]^{1-\delta_i} \\ & + r_i z_i \log \left[\frac{\lambda_R \lambda_1}{\lambda_1 - \lambda_R} e^{-\lambda_R u_i} + \frac{\lambda_R \lambda_1}{\lambda_R - \lambda_1} e^{-\lambda_1 u_i} \right]^{\delta_i} \left[\frac{\lambda_1}{\lambda_1 - \lambda_R} e^{-\lambda_R u_i} + \frac{\lambda_R}{\lambda_R - \lambda_1} e^{-\lambda_1 u_i} \right]^{1-\delta_i} \\ & + r_i (1 - z_i) \log \left[\frac{\lambda_R \lambda_2}{\lambda_2 - \lambda_R} e^{-\lambda_R u_i} + \frac{\lambda_R \lambda_2}{\lambda_R - \lambda_2} e^{-\lambda_2 u_i} \right]^{\delta_i} \left[\frac{\lambda_2}{\lambda_2 - \lambda_R} e^{-\lambda_R u_i} + \frac{\lambda_R}{\lambda_R - \lambda_2} e^{-\lambda_2 u_i} \right]^{1-\delta_i} \}. \end{aligned} \tag{3.24}$$

Since the full likelihood factorizes into two parts, each part can be maximized separately. The maximum likelihood estimates for $L_1(\pi; O)$ are given in (3.20) above. $L_2(\theta; O)$ can be maximized numerically since the estimates of the parameters from the convolution do not have close form solutions. Assuming an exponential distribution for $k = 1$ leads to

$$\hat{S}_{A_1 B_1}(u) = (1 - \hat{\pi}_r) e^{-\hat{\lambda}_0 u} + \hat{\pi}_r \left(\frac{\hat{\lambda}_1}{\hat{\lambda}_1 - \hat{\lambda}_R} e^{-\hat{\lambda}_R u} + \frac{\hat{\lambda}_R}{\hat{\lambda}_R - \hat{\lambda}_1} e^{-\hat{\lambda}_1 u} \right)$$

3.2.3 Large sample properties

Consider the case when $k = 1$, that is when estimating survival curve for treatment policy $A_1 B_1$

$$\hat{S}_{A_1 B_1}(u) = (1 - \hat{\pi}_r) \hat{S}_0(u) + \hat{\pi}_r \hat{S}_1(u), \quad \text{for } u \in [0, \tau].$$

Let $\hat{\phi} = (\hat{\pi}_r, \hat{\theta})$ and $G(u)$ denote the vector of partial derivatives with respect to each parameter in $\phi = (\pi_r, \theta)$. Also define $V = var(\hat{\phi})$ to be the variance-covariance matrix for the MLEs. Then by the delta method, we have that

$$\hat{S}_{A_1 B_1}(u)) \; \dot{\sim} \; N(S_{A_1 B_1}(u), \Sigma(u)) \tag{3.25}$$

where $\Sigma(u) = G(u)VG(u)^T$. We estimate $\Sigma(u)$ by replacing $\phi = (\pi_r, \theta)$ with $\hat{\phi} = (\hat{\pi}_r, \hat{\theta})$, this leads to $\hat{\Sigma} = \hat{G}\hat{V}\hat{G}^T$, where $\hat{V}$ is the estimated variance-covariance matrix of $\hat{\phi}$.

Example: Exponential model

Using the delta method, we compute the variance of $\hat{S}_{A_1B_1}$, when exponential distributions are assumed. Taking partial derivatives with respect to the parameters, we get

$$\begin{aligned}
d_1 &= \frac{\partial S_{A_1B_1}(u)}{\partial \pi_r} \\
&= -e^{-\lambda_0 u} + \left(\frac{\lambda_1}{\hat{\lambda}_1 - \lambda_R} e^{-\lambda_R u} + \frac{\lambda_R}{\lambda_R - \lambda_1} e^{-\lambda_1 u} \right), \\
d_2 &= \frac{\partial S_{A_1B_1}(u)}{\partial \lambda_0} \\
&= -u(1-\pi_r)e^{-\lambda_0 u}, \\
d_3 &= \frac{\partial S_{A_1B_1}(u)}{\partial \lambda_R} \\
&= \frac{\lambda_1}{(\lambda_1 - \lambda_R)^2} e^{-\lambda_R u} - \frac{\lambda_1 u}{\lambda_1 - \lambda_R} e^{-\lambda_R u} - \frac{\lambda_1}{(\lambda_R - \lambda_1)^2} e^{-\lambda_1 u}, \\
d_3 &= \frac{\partial S_{A_1B_1}(u)}{\partial \lambda_1} \\
&= \frac{\lambda_R}{(\lambda_R - \lambda_1)^2} e^{-\lambda_1 u} - \frac{\lambda_R}{(\lambda_1 - \lambda_R)^2} e^{-\lambda_R u} - \frac{\lambda_R u}{\lambda_R - \lambda_1} e^{-\lambda_1 u}.
\end{aligned}$$

Now, given $G = \begin{pmatrix} d_1 & d_2 & d_3 & d_4 \end{pmatrix}$, we obtain

$$\Sigma = G \begin{pmatrix} var(\hat{\pi}_r) & 0 & 0 & 0 \\ 0 & var(\hat{\lambda}_0) & 0 & 0 \\ 0 & 0 & var(\hat{\lambda}_R) & cov(\hat{\lambda}_R, \hat{\lambda}_1) \\ 0 & 0 & cov(\hat{\lambda}_1, \hat{\lambda}_R) & var(\hat{\lambda}_1) \end{pmatrix} G^T. \tag{3.26}$$

We plug-in $\hat{\phi}$ to obtain $\hat{G} = \begin{pmatrix} \hat{d}_1 & \hat{d}_2 & \hat{d}_3 & \hat{d}_4 \end{pmatrix}$. $V = var(\hat{\phi})$ is estimated by the observed Fisher information matrix.

3.3 Simulation study

To study the performance of the proposed estimator, a simulation study was conducted and comparison with other estimators was done. The generation of the datasets was done following a two-stage SMART design with two first stage treatments and two second stage treatments. We focused on data from A_1 as data from A_1 and A_2 are independent. All simulation were done in R.

Different simulation scenarios were considered with different response rates. R_i was taken to be a Bernoulli distribution with $P(R_i = 1) = \pi_r$, and $\pi_r \in (0.5, 0.7)$ so as to achieve 50%, and 70% of individuals responding to the first stage intervention. T_{0i} was generated from an exponential distribution with mean of 3 years for those with $R_i = 0$. The second stage indicator was generated from a Bernoulli distribution with $P(Z_i = 1) = \pi_z$, and π_z was set to be 0.5 in all simulations. We generated T_i^R from an exponential distribution with a mean of 5 years for the responders to the first stage treatment. For those with $Z_i = 1$, T_{1i}^* was generated from an exponential distribution with a mean of 7 years and T_{2i}^* was generated from an exponential distribution with a mean of 8 years. The observed survival time, T_i, was obtained using equation (1.2). The right censoring time,C_i,, was generated from a uniform distribution, $U(0, v)$, such that 20% and 40% of the sample were censored. Finally, the observed time was defined as $U_i = \min(T_i, C_i)$.

The DTR package was used for simulating the WRSE and the LDT estimator (Tang and Melguizo, 2005). For our estimator, an ad-hoc R function was written and maximized using optim function in R. For purely abbreviation purposes, we denote our parametric approach for time -varying SMART designs as TVS. The simulation study is done similar to the one in Chapter 1.

TABLE 3.1: Simulation Results

			TVS				**WRSE**					**LDT**			
π_r	t	True	$\hat{S}_1(u)$	SE	Bias	CP	$\hat{S}_1(u)$	SE	Bias	CP	RE	$\hat{S}_1(u)$	SE	Bias	CP
							$n = 100$	c=20%							
0.5	1	0.851	0.859	0.021	0.01	95.2	0.855	0.035	0.00	95.7	0.370	0.851	0.038	0.00	95.6
	3	0.638	0.642	0.041	0.00	95.4	0.648	0.051	0.01	94.7	0.652	0.640	0.057	0.00	95.1
	6	0.434	0.439	0.048	0.01	94.2	0.453	0.057	0.02	93.7	0.715	0.439	0.067	0.01	93.4
	8	0.340	0.347	0.049	0.01	94.6	0.367	0.058	0.03	92.2	0.736	0.350	0.068	0.01	93.1
	12	0.211	0.223	0.047	0.01	94.6	0.241	0.055	0.03	92.5	0.778	0.220	0.064	0.01	89.4
							$n = 300$	c=20%							
0.5	1	0.851	0.854	0.013	0.00	96.1	0.852	0.021	0.00	95.0	0.370	0.850	0.022	0.00	95.6
	3	0.638	0.637	0.024	0.00	95.9	0.643	0.030	0.01	95.7	0.637	0.639	0.034	0.00	96.6
	6	0.434	0.432	0.028	0.00	95.9	0.447	0.034	0.01	94.2	0.707	0.441	0.039	0.01	94.8
	8	0.340	0.339	0.029	0.00	95.0	0.358	0.034	0.02	93.9	0.723	0.350	0.040	0.01	94.5
	12	0.211	0.213	0.027	0.00	94.6	0.232	0.032	0.02	90.5	0.728	0.223	0.038	0.01	92.7
							$n = 100$	c=40%							
0.5	1	0.851	0.860	0.022	0.01	95.3	0.855	0.036	0.00	95.6	0.385	0.838	0.038	0.01	89.9
	3	0.638	0.643	0.042	0.01	94.6	0.648	0.052	0.01	94.6	0.672	0.607	0.060	0.03	86.0
	6	0.434	0.444	0.051	0.01	94.7	0.452	0.060	0.02	92.9	0.737	0.387	0.070	0.05	77.4
	8	0.340	0.355	0.054	0.02	95.2	0.368	0.062	0.03	93.1	0.769	0.292	0.071	0.05	75.3
	12	0.211	0.238	0.055	0.03	95.1	0.243	0.061	0.03	90.7	0.847	0.150	0.061	0.06	62.1
							$n = 300$	c=40%							
0.5	1	0.851	0.854	0.013	0.00	94.2	0.853	0.021	0.00	95.8	0.387	0.839	0.022	0.01	86.2
	3	0.638	0.637	0.025	0.00	94.9	0.643	0.030	0.01	94.0	0.653	0.607	0.036	0.03	78.5
	6	0.434	0.433	0.029	0.00	95.1	0.447	0.035	0.01	94.1	0.718	0.392	0.043	0.04	72.9
	8	0.340	0.342	0.029	0.00	94.6	0.358	0.037	0.02	93.3	0.729	0.293	0.044	0.05	68.6
	12	0.211	0.219	0.030	0.01	94.8	0.235	0.037	0.02	92.1	0.731	0.158	0.041	0.05	59.6
							$n = 100$	c=20%							
0.7	1	0.906	0.912	0.018	0.01	94.8	0.908	0.029	0.00	95.4	0.374	0.906	0.031	0.00	95.6
	3	0.746	0.747	0.037	0.00	94.9	0.755	0.048	0.01	94.7	0.623	0.748	0.052	0.00	95.0
	6	0.553	0.552	0.050	0.00	94.1	0.573	0.060	0.02	93.3	0.710	0.562	0.068	0.01	93.3
	8	0.449	0.450	0.053	0.00	94.6	0.476	0.063	0.03	92.3	0.739	0.461	0.071	0.01	92.4
	12	0.288	0.296	0.054	0.01	94.0	0.322	0.062	0.03	92.5	0.769	0.302	0.071	0.01	90.0
							$n = 300$	c=20%							
0.7	1	0.906	0.908	0.010	0.00	94.4	0.907	0.017	0.00	94.4	0.372	0.906	0.018	0.00	94.4
	3	0.746	0.746	0.022	0.00	94.8	0.754	0.028	0.01	93.3	0.596	0.751	0.031	0.01	93.3
	6	0.553	0.551	0.029	0.00	95.5	0.573	0.035	0.02	92.2	0.697	0.567	0.039	0.01	92.2
	8	0.449	0.448	0.031	0.00	95.0	0.475	0.037	0.03	91.8	0.726	0.468	0.042	0.02	91.8
	12	0.288	0.290	0.031	0.00	94.4	0.320	0.037	0.03	89.6	0.717	0.310	0.042	0.02	89.6
							$n = 100$	c=40%							
0.7	1	0.906	0.912	0.018	0.01	94.2	0.908	0.030	0.00	95.1	0.392	0.892	0.032	0.01	87.6
	3	0.746	0.747	0.039	0.00	94.2	0.753	0.049	0.01	95.0	0.651	0.712	0.055	0.03	83.4
	6	0.553	0.556	0.053	0.00	94.2	0.576	0.063	0.02	92.8	0.744	0.506	0.072	0.05	77.0
	8	0.449	0.456	0.059	0.01	94.0	0.481	0.067	0.03	91.6	0.770	0.395	0.076	0.05	73.8
	12	0.288	0.310	0.062	0.02	94.2	0.328	0.071	0.04	91.0	0.806	0.215	0.072	0.07	61.8
							$n = 300$	c=40%							
0.7	1	0.906	0.908	0.011	0.00	95.2	0.908	0.017	0.00	95.0	0.391	0.895	0.019	0.01	86.3
	3	0.746	0.745	0.022	0.00	95.2	0.753	0.029	0.01	94.0	0.626	0.719	0.033	0.03	80.3
	6	0.553	0.552	0.031	0.00	94.1	0.572	0.037	0.01	93.4	0.723	0.513	0.043	0.04	75.4
	8	0.449	0.450	0.034	0.00	93.2	0.474	0.040	0.03	92.0	0.747	0.401	0.047	0.5	72.3
	12	0.288	0.296	0.036	0.01	94.2	0.319	0.043	0.03	91.4	0.738	0.224	0.047	0.06	57.0

Table 3.1 shows the results of the simulation study. The results for our estimator are given under the TVS columns. We report the standard errors, absolute bias, and 95% coverage probabilities (CP) for the three estimators for treatment policy A_1B_1. Relative efficiency (RE) is also reported between our parametric estimator and the WRSE. The relative efficiency is calculated as sample variance of our estimator divided by the sample variance of the WRSE for estimating the survival function. Guo and Tsiatis (2005) established that the WRSE is more efficient as compared to the LDT

estimator, for this reason, we only computed the relative efficiency of our estimator and the WRSE. Two different censoring and response rates are considered.

The results of this simulation study show that our estimator is more precise compared to its nonparametric counterparts. This is shown by the small standard errors across all the simulation scenarios. The LDT estimator has the largest standard errors among the estimators. Our estimator is more efficient than the other two estimators. This result is not surprising. Inferences based on parametric distributions is more precise provided the parametric assumptions are valid (Collett, 2015). The coverage probabilities of our method are close to the nominal level, the same applies to the WRSE. The coverage probabilities of the LDT estimator are highly affected by the change in censoring rates. In cases where the censoring rate is high, that is, 40% the coverage probabilities are way below the desired nominal level.

In terms of biasedness, all the three methods performed fairly well with the exception of the LDT estimator in the case of 40% censoring. Increasing the censoring rate from 20% to 40% for the LDT estimator leads to an increase in bias. There is, however, a minimal increase in bias for the other two estimators when the censoring rate was increased. The bias vanished with increase in sample size as expected. Our parametric estimator has the least bias among the three methods, and when the sample size is 300, the bias of our method diminishes. Changing the response rates changes the survival estimates. In general all the three methods yield similar survival estimates. The differences in the survival estimates is profound for the LDT when the censoring rate is 40%. With a lower censoring rate, the survival estimates from the three methods are mostly similar.

3.4 Convolutions by numerical methods

The methodology developed in this chapter is not restricted to the exponential distribution. The exponential distribution is used to illustrate this methodology because of the availability of closed form solution for the convolution integral. Convolution integrals for other parametric distributions are hard to solve. The methodology can be generalized to other distributions using either numerical approaches (Ruckdeschel and Kohl, 2010) or approximations to the convolution integral (Bessate and El Bouanani, 2016). In statistics, convolution of probability distributions is a standard problem. Fast Fourier Transforms (FFT) have been used in approximating convolutions numerically.

3.4.1 FFTs and convolutions

The Fourier transform of a convolution of two functions is the product of the Fourier transforms of the functions. Let f and g be two functions with convolution $f * g$, and let $\mathcal{F}$ denote the Fourier transform operator such that $\mathcal{F}(f)$ and $\mathcal{F}(g)$ are the Fourier transforms of f and g respectively. The Fourier transform of the convolution is then

$$\mathcal{F}(f * g) = \mathcal{F}(f) \cdot \mathcal{F}(g). \tag{3.27}$$

Using the inverse Fourier transform, we can write

$$f * g = \mathcal{F}^{-1}\{\mathcal{F}(f) \cdot \mathcal{F}(g)\}. \tag{3.28}$$

This result also holds for Laplace transforms. For numerical computation of convolutions, the FFT is used for efficiency (Brigham and Morrow, 1967).

Algorithm

Let $f = (f_0, f_1, ..., f_{m-1})$ and $g = (g_0, g_1, ..., g_{k-1})$ be two discrete probability vectors, the convolution can be evaluated using the following algorithm:

- Pad the given vectors f and g with zeroes so that each has length $n \geq m + k$.
- Apply to each vector the FFT, $\tilde{f} = FFT(f)$ and $\tilde{g} = FFT(g)$.
- Compute the inner product for the two vectors, $\tilde{h} = \tilde{f} \cdot \tilde{g}$.
- Lastly, apply the inverse function of the FFT to $\tilde{h}$ to obtain the probability vector as a convolution of f and g.

3.4.2 Implementation

The 'distr' package in R computes convolutions of distributions using the FFT (Ruckdeschel and Kohl, 2010). In this package, the convolutions can be computed for almost any arbitrary univariate distribution. The distributions can either be continuous or discrete. The convolution operator "+" returns a distribution object consisting not only of either a cumulative distribution function (cdf) or a density function but automatically all four constitutive functions, that is, cdf, density, quantile function and a random number generator. For example, if X_1 and X_2 are two independent Weibull random variables, $Y = X_1 + X_2$ can be computed in the distr package.

```
> X1 <- Weibull(shape=1.2, scale=1.5)
> X2 <- Weibull(shape=2, scale=2.5)
> Y <- convpow(X1 + X2, 1)
```

```
> d(Y)(0.8)
[1] 0.04505478
> q(Y)(1/3)
[1] 2.762783
> r(Y)(5)
[1] 2.869811 6.237388 2.506635 4.457062 1.943053
```

3.5 Application: CALGB 19808 study

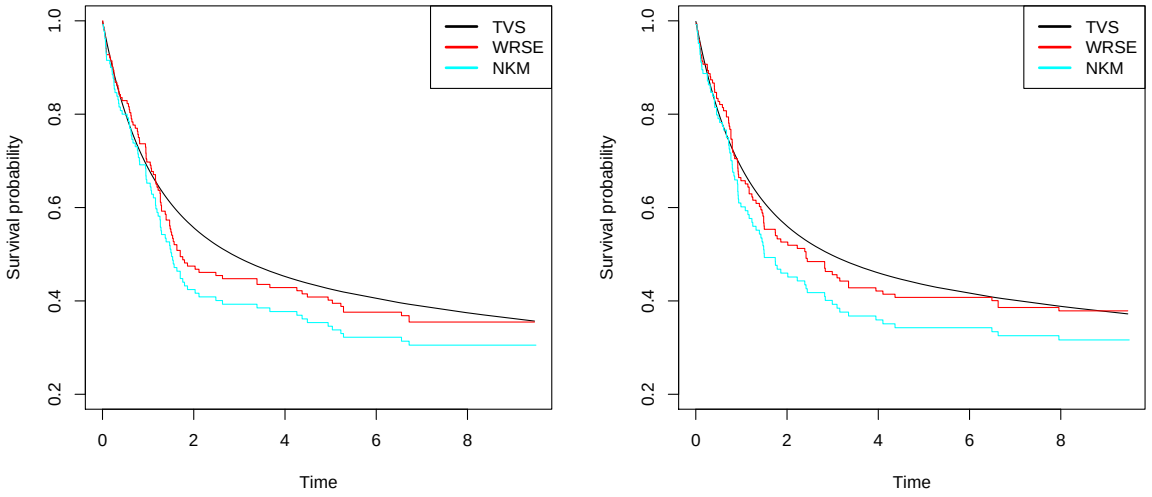
We apply our methodology to the CALGB 19808 study. Table 3.2 summarizes the results of the application.

Table 3.2: Application results

t	Policy	$\hat{S}(t)^{exp}_{TVS}$	SE^{exp}	$\hat{S}(t)^{gom}_{TVS}$	SE^{gom}	$\hat{S}(t)_{WRSE}$	SE
0.03	ADEP-rIL-2	0.9952	0.0006	0.9845	0.0026	0.9868	0.0092
	ADEP-OBS	0.9952	0.0006	0.9844	0.0027	0.9868	0.0092
	ADE-rIL-2	0.9953	0.0006	0.9847	0.0027	0.9801	0.0114
	ADE-OBS	0.9953	0.0006	0.9847	0.0030	0.9801	0.0114
1.3	ADEP-rIL-2	0.8114	0.0186	0.6356	0.0402	0.6302	0.0407
	ADEP-OBS	0.8007	0.0188	0.6249	0.0429	0.5964	0.0436
	ADE-rIL-2	0.8146	0.0188	0.6427	0.0430	0.6157	0.0433
	ADE-OBS	0.8034	0.0189	0.6315	0.0420	0.6210	0.0422
4.1	ADEP-rIL-2	0.5233	0.0352	0.4422	0.0438	0.4286	0.0449
	ADEP-OBS	0.4890	0.0334	0.4079	0.0422	0.3815	0.0463
	ADE-rIL-2	0.5337	0.0356	0.4571	0.0419	0.4144	0.0471
	ADE-OBS	0.4988	0.0338	0.4221	0.0446	0.3624	0.0452
8.0	ADEP-rIL-2	0.3186	0.0389	0.3676	0.0262	0.3547	0.0445
	ADEP-OBS	0.2663	0.0325	0.3153	0.0435	0.3199	0.0452
	ADE-rIL-2	0.3344	0.0389	0.3888	0.0331	0.3787	0.0470
	ADE-OBS	0.2796	0.0329	0.3341	0.0311	0.3407	0.0450

Table 3.2 shows the results of fitting our method to the CALGB 19808 study. This analysis is based on the overall survival. For the first component in the survival mixture model (S_0), we assumed either the exponential or the Gompertz distributions. Under the columns $\hat{S}(t)^{exp}_{TVS}$ and $\hat{S}(t)^{gom}_{TVS}$, we report the survival estimates when the exponential or the Gompertz distribution is used for the non-responders. The second component, $(T^R_i + T^*_{ki})$, is the convolution of exponential distributions. The results when a Gompertz

distribution was used provide a better fit with similar estimates to the WRSE. This is not surprising as the Gompertz distribution provided a better fit to the data for non-responders from the comparisons in Figure 2.2. The fit with an exponential distribution gives a poor fit. It tends to overestimate the survival probability in the middle of the curve and the discrepancy is profound.



(I) ADEP-rIL-2 (II) ADE-rIL-2

FIGURE 3.1: Survival curves using the Gompertz distribution for non-responders.

Figure 3.1 shows the survival curves for two treatment policies when the Gompertz distribution is used for the non-responders in the survival mixture model. The fit is plotted together with the curve from the WRSE and the naive Kaplan-Meier estimator (NKM). The survival curves from the two methods are similar except in the middle part of the curves where the parametric method tends to overestimate the survival probability. The difference is minimal except in the middle of the curve. In both cases, the naive Kaplan-Meier underestimates the survival distributions, and the differences are quite large after 2 years. This result is consistent with the results from Guo and Tsiatis (2005), where the naive Kaplan-Meier method was found to be biased in estimating the survival curve.

3.6 Conclusion

We can differentiate between two types of SMART designs. In some SMART designs, the response is measured at one time point. In some other SMART designs, response is

measured at different time points in the first stage. The time to response then differs from patient to patient. This makes the observed survival times differ in these two types of designs. In the latter, the observed survival times are a sum of two random survival times for the responders. This makes it to be theoretically flawed to just assume a single survival model for the sum. The density of a sum of two random variable is always given by a convolution. In this chapter, we have proposed to use convolution based density function in modeling the total times for responders. Maximum likelihood estimation was used and the results are compared to the non-parametric estimates from the WRSE. The proposed approach is not restricted to only convolutions of the exponential distribution but can be generalized to other distributions using numerical methods based on the discrete Fourier transforms or other approximations. The distr package provides a platform where such probability densities can be computed.

We note that the way the survival time is defined in a standard SMART design makes it easier for the parametric analysis to be conducted as it avoids the use of convolutions. The way the survival time is defined for responders in time-varying SMART designs poses a challenge in the analysis.

Chapter 4

Weighted Lin and Xu test for two-stage randomization designs

4.1 Motivating example

In Figure 2.5, we showed the survival curves for the four treatment strategies in the CALGB 19808 study. To estimate the survival curves consistently, the WRSE of Guo and Tsiatis (2005) was used. These curves are weighted using the inverse probabilities of being in that particular treatment strategy of interest. We are interested in comparing overall survival between the treatment strategies. The survival curves cross at some time points. In cases where the survival curves cross, standard comparison techniques could lead to misleading conclusions.

In comparing the survival curves, we want to find the sequence that yields better long-term survival. To accomplish this, one may select a single long-term time point and then compare the survival estimates between the treatment strategies. This approach, however, may not work well since results can be sensitive to the time point selected (Klein et al., 2007). Another approach is to use the weighted log-rank test (Kidwell and Wahed, 2013). Even with carefully chosen weight functions there is no guarantee that the weighted log-rank test would perform well when the survival curves cross. Kidwell and Wahed (2013) proposed a modification of the log-rank test to be used in comparing treatment policies in two-stage designs. It is not guaranteed that the weighted log-rank test would work well in cases of crossing survival curves since none of their simulation scenarios addressed directly the case where the survival curves cross. To this end, we propose the weighted Lin and Xu test. The Lin and Xu (2010) test is based on the

absolute differences between the two survival curves and is robust in cases when the proportional hazards assumption is violated.

4.2 Weighted log-rank test

Kidwell and Wahed (2013) proposed the weighted log-rank test for comparing treatment policies in two-stage SMART designs. Consider testing the hypothesis $H_0 : \Lambda_{11}(t) - \Lambda_{12}(t)$. The treatment strategies A_1B_1 and A_1B_2 are shared path strategies. The group of non-responders is the same in both treatment strategies hence they are not independent. In the case where there was no second stage randomization, that is, if patients are randomized to follow either A_1B_1 or A_1B_2, in such an instance, data from patients receiving A_1B_1 could be regarded as independent of data from patients receiving A_1B_2. In such a case, the unweighted log-rank test for comparing two independent samples is employed,

$$Z_n(t) = \int_0^t \frac{Y_{11}(s)Y_{12}(s)}{Y_{11}(s) + Y_{12}(s)} \left\{ \frac{dN_{11}(s)}{Y_{11}(s)} - \frac{dN12(s)}{Y_{12}(s(} \right\},$$

where $N_{1ki}(s) = I(U_{1ki} \leq s, \Delta_{1ki} = 1)$, $Y_{1ki}(s) = I(U_{1ki} \geq s)$, $N_{1k}(s) = \sum_{i=1}^n N_{1ki}$, $Y_{1k}(s) = \sum_{i=1}^n Y_{1ki}(s)$ for $k = 1, 2$. When the null hypothesis is true, $n^{1/2}Z_n(t)$ is asymptotically normally distributed with mean zero and variance that can estimated consistently from the observed event times.

On the other hand, in a two-stage SMART design, patients are randomized to the second stage treatments. Assuming independence between A_1B_1 and A_1B_2 cannot be done. To account for the second stage randomization, a weighted version of standard log-rank (SLR) test should be used and to account for the dependence between A_1B_1 and A_1B_2 , a modification of the SLR should be made.

4.2.1 Weighted two-sample statistic

To develop the two-sample statistic, time dependent weights similar to those in Guo and Tsiatis (2005) are used. Let $W_{11}(s) = (X_i/\phi)\{1 - R_i(s) + R_i(s)Z_i/\pi_z\}$ be the weight function given to patient i at time s for the purposes of estimating quantities related to treatment policy A_1B_1. $R_i(s) = R_iI(T_i^R \leq s)$ such that if at time s patient i has responded then $R_i(s) = 1$, and zero otherwise; π_z is the probability of being randomized to B_1, and ϕ is the probability of being randomized to A_1. A similar weight is defined for A_1B_2 where now $W_{12}(s) = (X_i/\phi)\{1 - R_i(s) + R_i(s)(1 - Z_i)/(1 - \pi_z)\}$. For non-responders, $W_{11}(s) = W_{12}(s) = 1/\phi$, non-responders are consistent with both

treatment policies. For responders assigned to A_1B_1, $W_{11}(s) = 1/\phi\pi_z$ and $W_{12} = 0$. If the patient has responded and is assigned to B_2 at time s then, $W_{11i}(s) = 0$ and $W_{12i}(s) = 1/\phi(1-\pi_z)$. With this weight function, the weighted log-rank test statistic for testing $H_0 : \Lambda_{11}(t) - \Lambda_{12}(t)$ is

$$Z_n^w(t) = \int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{\bar{Y}_{11}(s) + \bar{Y}_{12}(s)} \left\{ \frac{d\bar{N}_{11}(s)}{\bar{Y}_{11}(s)} - \frac{d\bar{N}12(s)}{\bar{Y}_{12}(s(} \right\},$$

where $\bar{Y}_{jk}(s) = \sum_{i=1}^n W_{jki} Y_i(s)$, and $\bar{N}_{jk}(s) = \sum_{i=1}^n W_{jki} N_i(s)$.

When H_0 is true, the weighted log-rank statistic has expectation zero since with respect to $F(t) = \sigma\{X_i, R_i(s), R_i(s)Z_i, I(C_i \leq s), N_i(s), i = 1, ..., n; j = 1, 2; 0 \leq s \leq t\}$, the term $\{\bar{Y}_{11}(s)\bar{Y}_{12}(s)/\bar{Y}_{11}(s) + \bar{Y}_{12}(s)\}$ is predictable. The variance is modified to account for the correlation between the two treatment policies. A consistent variance estimator of $n^{1/2}Z_n^w(t)$ is given by

$$\begin{aligned} \hat{\sigma}^2(t) = n^{-1} \int_0^t \frac{\bar{Y}_{12}^2(s)\sum_{i=1}^n W_{11i}^2(s)Y_i(s) + \bar{Y}_{11}^2(s)\sum_{i=1}^n W_{12i}^2(s)Y_i(s)}{(\bar{Y}_{11}(s) + \bar{Y}_{12}(s))^2} \left\{ \frac{dN_{1.}(s)}{Y_{1.}(s)} \right\} \\ - 2(n\phi^2)^{-1} \int_0^t \frac{\bar{Y}_{11}(s)\bar{Y}_{12}(s)}{(\bar{Y}_{11}(s) + \bar{Y}_{12}(s))^2} \left\{ Y_1^{NR}(s) \frac{dN_1^{NR}(s)}{Y_1^{NR}(s)} \right\}, \end{aligned} \tag{4.1}$$

where $Y_j^{NR} = \sum_{i=1}^n I(X_i = 2-j)(1 - R_i(s))Y_i(s)$ is the number of individuals who have yet to respond to treatment A_j and are at risk at time s, $Y_{j.}(s) = \sum_{i=1}^n I(X_i = 2-j)Y_i(s)$ is the number of individuals with initial treatment A_j and are at risk at time s, $N_j^{NR} = \sum_{i=1}^n I(X_i = 2-j)(1 - R_i(s))N_i(s)$, and, $N_{j.}(s) = \sum_{i=1}^n I(X_i = 2-j)N_i(s)$. A table of the description of the notation is found in Kidwell and Wahed (2013). For testing equality between two independent path treatment policies one simply ignores the covariance term in the variance formula above.

4.3 Lin and Xu test

The SLR test has optimal power in detecting differences between two survival distributions if the assumption of proportional hazards is not violated. In clinical trials with survival endpoints, survival curves may cross at some point and this complicates the analysis as the usual SLR test may not work well. Alternatives to the SLR test includes weighted versions of the same, Kolmogorov-Smirnov test and many others. The Kolmogorov-Smirnov test has better power than the Wilcoxon test, a weighted version of SLR, when the survival curves cross but there is no guarantee that the Kolmogorov-Smirnov test is always better in the case of crossing survival curves. Lin and Xu (2010)

proposed a test for the equality of survival distributions that is robust in instances where the assumption of proportional hazards is violated. This test is based on the absolute difference between the survival curves.

4.3.1 Notation and assumptions

Consider a comparison of survival experience between two treatment groups. Denote by $U = min(T, C)$, where T is the survival time and C is the censoring time. Let $\delta = I(T \leq C)$ be an indicator for an event, that is, $\delta = 1$ if an event is observed and 0 otherwise. Let G be the group indicator, $G = 1$ for group I and $G = 2$ for group II. Given G, we assume that C is independent of T. Denote by N_G the sample size in G. For $G = 1, 2$, let $U_{G(i)}$, $i = 1, 2...N_g$ be the ordered survival times in each group with δ_{Gi} being the corresponding survival status. Denote the pooled distinct event times in the two samples as $t_1 < t_2 < ... < t_k$. Let d_{G_j} and n_{G_j} be the number of observed events and the number at risk at time t_j in the two groups. To estimate the survival distribution at time t, $S_G(t)$, the Kaplan-Meier estimator is used:

$$\hat{S}_G(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_{G_j}}{n_{G_j}}\right).$$

For the two survival curves, the observed absolute difference is given by

$$\begin{aligned} \Delta &= \int_0^\tau |\hat{S}_1(t) - \hat{S}_2(t)| dt \\ &= \sum_{j:t_j < \tau} |\hat{S}_1(t_j) - \hat{S}_2(t_j)|(t_{j+1} - t_j), \end{aligned} \tag{4.2}$$

where τ is the largest time point by which areas under the curves can be calculated for both groups based on data available.

We define τ for three cases: $\tau = \min_G(U_{G(N_G)})$ if the last two points in the two groups are both censored; $\tau = \max_G(U_{G(N_G)}(1 - \delta_{G(N_G)}))$ if, in one group, the last time point is an event and in the other group it is a censored observation, and finally $\tau = \max_G(U_{G(N_G)})$ if the last time points in both groups are events. To estimate the variance of $\hat{S}_G(t)$, the Greenwood formula is used

$$\hat{\sigma}^2_{S_G} = \hat{S}^2_G(t) \sum_{j:t_j \leq t} \frac{d_{G_j}}{n_{Gj}(n_{Gj} - d_{Gj})}. \tag{4.3}$$

4.3.2 Test statistic

If Z is a standard normal random variable then the density of $|Z|$ is given by

$$f_{|Z|}(z) = \begin{cases} \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, & \text{if } z \geq 0 \\ 0, & \text{if } z < 0. \end{cases}$$

The expectation of $|Z|$ is

$$\begin{aligned} E[|Z|] &= \sqrt{\frac{2}{\pi}} \int_0^\infty z e^{-\frac{1}{2}z^2} \\ &= \sqrt{\frac{2}{\pi}}, \end{aligned}$$

and

$$\begin{aligned} E[|Z|]^2 &= \sqrt{\frac{2}{\pi}} \int_0^\infty z^2 e^{-\frac{1}{2}z^2} \\ &= 1 \end{aligned}$$

hence

$$\text{var}|Z| = 1 - \frac{2}{\pi}.$$

Consider $H_0 : S_1(t) - S_2(t) = 0$, then if H_0 is true $S_1(t) - S_1(t)$ is approximately normally distributed with mean 0 and variance $[\sigma^2_{S_1}(t) + \sigma^2_{S_2}(t)]$. Then

$$\hat{E}[|\hat{S}_1(t) - \hat{S}_2(t)|] \doteq \left\{2/\pi[\hat{\sigma}^2_{S_1}(t) + \hat{\sigma}^2_{S_2}(t)]\right\}^{\frac{1}{2}},$$

and

$$\widehat{\text{Var}}[|\hat{S}_1(t) - \hat{S}_2(t)|] \doteq (1 - \frac{2}{\pi})[\hat{\sigma}^2_{S_1}(t) + \hat{\sigma}^2_{S_2}(t)].$$

Using the normal approximation by the Greenwood's formula we get

$$\hat{E}[\Delta] \doteq \sum_{j:t_j<\tau} \left\{2/\pi[\hat{\sigma}^2_{S_1}(t_j) + \hat{\sigma}^2_{S_2}(t_j)]\right\}^{\frac{1}{2}} (t_{j+1} - t_j). \tag{4.4}$$

The variance is estimated by

$$\begin{aligned}
\widehat{\text{Var}}(\Delta) &= Var\left(\sum_{j:t_j<\tau} |\hat{S}_1(t_j) - \hat{S}_2(t_j)|(t_{j+1} - t_j)\right) \\
&= \sum_{j:t_j<\tau} (t_{j+1} - t_j)^2 Var\left\{|\hat{S}_1(t_j) - \hat{S}_2(t_j)|\right\} \\
&+ \sum_{j<j':t_j,t_{j'}<\tau} 2\text{Cov}\left\{|\hat{S}_1(t_j) - \hat{S}_2(t_j)|(t_{j+1} - t_j), |\hat{S}_1(t_{j'}) - \hat{S}_2(t_{j'})|(t_{j'+1} - t_{j'})\right\} \\
&\doteq \sum_{j:t_j<\tau} (t_{j+1} - t_j)^2 (1 - 2/\pi)[\hat{\sigma}^2_{S_1}(t_j) + \hat{\sigma}^2_{S_2}(t)] \\
&+ \sum_{j:t_j<\tau} 2(t_{j+1} - t_j)(t_{j'+1} - t_{j'})\text{Cov}\left\{|\hat{S}_1(t_j) - \hat{S}_2(t_j)|, |\hat{S}_1(t_{j'}) - \hat{S}_2(t_{j'})|\right\} \\
&= \sum_{j:t_j<\tau} (t_{j+1} - t_j)^2 (1 - 2/\pi)[\hat{\sigma}^2_{S_1}(t_j) + \hat{\sigma}^2_{S_2}(t)] \\
&+ \sum_{j:t_j<\tau} 2\rho_{j,j'}(t_{j+1} - t_j)(t_{j'+1} - t_{j'})(1 - 2/\pi)\left\{[\hat{\sigma}^2_{S_1}(t_j) + \hat{\sigma}^2_{S_2}(t_j)][\hat{\sigma}^2_{S_1}(t_{j'}) + \hat{\sigma}^2_{S_2}(t_{j'})]\right\}^{\frac{1}{2}},
\end{aligned}$$

where $\rho_{j,j'}$ is the correlation coefficient between $|\hat{S}_1(t_j) - \hat{S}_2(t_j)|$ and $|\hat{S}_1(t_{j'}) - \hat{S}_2(t_{j'})|$, $j \neq j'$.

The variance depends on the correlation coefficient and Lin and Xu (2010) suggested setting $\rho_{j,j'} = 0.5$ for all j and j'. With this choice, the variance of Δ can be estimated by

$$\begin{aligned}
\widehat{Var}(\Delta) &\doteq \sum_{j:t_j<\tau} (t_{j+1} - t_j)^2 (1 - 2/\pi)[\hat{\sigma}^2_{S_1}(t_j) + \hat{\sigma}^2_{S_2}(t)] \\
&+ \sum_{j:t_j<\tau} (t_{j+1} - t_j)(t_{j'+1} - t_{j'})(1 - 2/\pi) \\
&\times \left\{[\hat{\sigma}^2_{S_1}(t_j) + \hat{\sigma}^2_{S_2}(t_j)][\hat{\sigma}^2_{S_1}(t_{j'}) + \hat{\sigma}^2_{S_2}(t_{j'})]\right\}^{\frac{1}{2}}.
\end{aligned} \tag{4.5}$$

Lin and Xu (2010) proposed using the test statistic

$$\Delta^* = \frac{\Delta - \hat{E}(\Delta)}{\sqrt{\widehat{\text{Var}}(\Delta)}}, \tag{4.6}$$

which is asymptotically normally distributed with mean 0 and variance 1. Large values of Δ^* leads to rejection of the null hypothesis, that is, we reject H_0 if $|\Delta^*| > Z_{1-\alpha/2}$ where $Z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ cutoff point for the standard normal random variable.

4.4 Weighted Lin and Xu test

As is, the Lin and Xu test cannot be applied in testing for equality between two treatment policies in two-stage designs. We propose a way of adapting the Lin and Xu test to two-stage designs for comparison of separate-path treatment strategies then later we extend the methodology to shared-path treatment strategies. For the comparison of treatment strategies, we propose the weighted Lin and Xu test where the weighted version of the survival distributions are used. To obtain weighted survival estimates, inverse probability weights are used. Instead of using Kaplan-Meier estimates we use estimates of survival distributions given by the WRSE (Guo and Tsiatis, 2005). The WRSE is a natural extension of the Nelson-Aalen estimator and makes use of time dependent weights.

Consider estimating survival distributions for treatment policy A_1B_1, then

$$\hat{S}_{A_1B_1}(t) = \exp\left\{-\int_0^t \frac{\sum_{i=1}^n W_i(u)dN_i(u)}{\sum_{i=1}^n W_i(u)Y_i(u)}\right\}. \tag{4.7}$$

where $W_i(u)$ is the weight function depending on time u, $N_i(u)$ is a counting process and $Y_i(u)$ is the at risk process. The variance is given by

$$\widehat{\text{Var}}(S_{A_1B_1}(t)) = n^{-1}\{S_{A_1B_1}(t)\}^2\hat{\sigma}^2, \tag{4.8}$$

where

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n \left(\int_0^t \frac{W_i(u)\left[dN_i(u) - Y_i(u)\left\{\frac{\sum_{i=1}^n W_i(u)dN_i(u)}{\sum_{i=1}^n W_i(u)Y_i(u)}\right\}\right]}{n^{-1}\sum_{i=1}^n W_i(u)Y_i(u)}\right)^2$$

Using the weighted versions of the survival curves

$$\begin{aligned}\Delta^w &= \int_0^\tau |\hat{S}_1^w(t) - \hat{S}_2^w(t)|dt \\ &= \sum_{j:t_j<\tau} |\hat{S}_1^w(t_j) - \hat{S}_2^w(t_j)|(t_{j+1} - t_j),\end{aligned} \tag{4.9}$$

that is, we calculate the observed absolute difference of the areas between the two survival curves. $\hat{S}_1^w(t)$ and $\hat{S}_2^w(t)$ are estimated using the WRSE. Under H_0, $\hat{S}_1^w(t) - \hat{S}_2^w(t)$ has an approximate normal distribution with mean 0 and variance $[\sigma^2_{\hat{S}_1^w} + \sigma^2_{\hat{S}_2^w}]$.

Based on the WRSE variance formula above, $E(\Delta^w)$ can be estimated by

$$\hat{E}[\Delta^w] \doteq \sum_{j:t_j<\tau} \left\{2/\pi[\hat{\sigma}^2_{S_1^w}(t_j) + \hat{\sigma}^2_{S_2^w}(t_j)]\right\}^{\frac{1}{2}} (t_{j+1} - t_j). \tag{4.10}$$

The variance is

$$\begin{aligned}\widehat{\mathrm{Var}}(\Delta^w) = &\sum_{j:t_j<\tau} (t_{j+1} - t_j)^2(1 - 2/\pi)[\hat{\sigma}^2_{S_1^w}(t_j) + \hat{\sigma}^2_{S_2^w}(t)] \\ &+ \sum_{j:t_j<\tau} 2\rho_{j,j'}(t_{j+1} - t_j)(t_{j'+1} - t_{j'})(1 - 2/\pi) \\ &\times \left\{[\hat{\sigma}^2_{S_1^w}(t_j) + \hat{\sigma}^2_{S_2^w}(t_j)][\hat{\sigma}^2_{S_1^w}(t_{j'}) + \hat{\sigma}^2_{S_2^w}(t_{j'})]\right\}^{\frac{1}{2}},\end{aligned} \tag{4.11}$$

where $\rho_{j,j'}$ is the correlation coefficient between $|\hat{S}^w_1(t_j) - \hat{S}^w_2(t_j)|$ and $|\hat{S}^w_1(t_{j'}) - \hat{S}^w_2(t_{j'})|$, $j \neq j'$. Setting $\rho_{j,j'}$ at 0.5 leads to the following variance estimator

$$\begin{aligned}\widehat{Var}(\Delta^w) \doteq &\sum_{j:t_j<\tau} (t_{j+1} - t_j)^2(1 - 2/\pi)[\hat{\sigma}^2_{S_1^w}(t_j) + \hat{\sigma}^2_{S_2^w}(t_j)] \\ &+ \sum_{j:t_j<\tau} (t_{j+1} - t_j)(t_{j'+1} - t_{j'})(1 - 2/\pi) \\ &\times \left\{[\hat{\sigma}^2_{S_1^w}(t_j) + \hat{\sigma}^2_{S_2}(t_j)][\hat{\sigma}^2_{S_1^w}(t_{j'}) + \hat{\sigma}^2_{S_2^w}(t_{j'})]\right\}^{\frac{1}{2}}\end{aligned} \tag{4.12}$$

The proposed test statistic is

$$\Delta^{*w} = \frac{\Delta^w - \hat{E}(\Delta^w)}{\sqrt{\widehat{\mathrm{Var}}(\Delta^w)}}, \tag{4.13}$$

which is asymptotically normally distributed with mean 0 and variance 1. Large values of Δ^{*w} leads to rejection of the null hypothesis, that is, we reject H_0 if $|\Delta^*| > Z_{1-\alpha/2}$ where $Z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ cutoff point for the standard normal random variable.

4.5 Weighted Lin and Xu test for shared-path treatment strategies

In two-stage designs, not only do we compare separate-path treatment strategies but also shared-path treatment strategies. Shared-path treatment strategies are dependent. The dependence arise because non-responders are common to paths sharing the same first stage treatments. To use all data available in comparing two shared-path treatment

strategies, a modification of the weighted Lin and Xu should be made. Consider the treatment policies A_1B_1 and A_1B_2, the survival functions for these treatment policies, that is, $S_1^w(t)$ and $S_2^w(t)$ are dependent. Under H_0, $S_1^w(t) - S_2^w(t)$ is approximately normally distributed with mean 0 and variance $[\sigma^2_{S_1^w} + \sigma^2_{S_2^w} - 2\text{Cov}(S_1^w(t), S_2^w(t))]$. Using an approximation by the WRSE variance formula, $E(\Delta^w)$ can be estimated as

$$\hat{E}_s[\Delta^w] \doteq \sum_{j:t_j<\tau} \left\{2/\pi[\hat{\sigma}^2_{S_1^w}(t_j) + \hat{\sigma}^2_{S_2^w}(t_j) - 2\text{Cov}(\hat{S}_1^w(t_j), \hat{S}_2^w(t_j))]\right\}^{\frac{1}{2}} (t_{j+1} - t_j). \quad (4.14)$$

The variances can be estimated from the WRSE. To estimate the covariance term in (4.14), we appeal to the following theorem.

Covariance inequality 1. Suppose that X_1, X_2 are two real valued random variables such that $E[X_1X_2]$ and $E[X_1^2]$,$E[X_2^2]$ are all finite, then

$$\text{Cov}^2(X_1, X_2) \leq V[X_1]V[X_2]. \quad (4.15)$$

Proof. Write $\text{Cov}(X_1, X_2) = E\{(X_1 - \mu_1)(X_2 - \mu_2)\}$ where $\mu_i = E[X_i]$, $i = 1, 2$. Define $\tilde{U}_i = X_i - \mu_i$, $i = 1, 2$. By the Cauchy-Schwartz inequality

$$\text{Cov}^2(X_1, X_2) = E^2(\tilde{U}_1\tilde{U}_2) \leq E[\tilde{U}_1]E[\tilde{U}_2] = V[X_1]V[X_2].$$

□

The covariance term can be approximated using the theorem above, that is,

$$\text{Cov}(\hat{S}_1^w(t), \hat{S}_2^w(t)) \approx \hat{\sigma}_{S_1^w}(t_j)\hat{\sigma}_{S_2^w}(t_j).$$

The variance can be estimated by

$$\begin{aligned}\widehat{Var}_s(\Delta^w) \doteq & \sum_{j:t_j<\tau} (t_{j+1} - t_j)^2(1 - 2/\pi)[\hat{\sigma}^2_{S_1^w}(t_j) + \hat{\sigma}^2_{S_2^w}(t_j) - 2\hat{\sigma}_{S_1^w}(t_j)\hat{\sigma}_{S_2^w}(t_j)] \\ & + \sum_{j:t_j<\tau} (t_{j+1} - t_j)(t_{j'+1} - t_{j'})(1 - 2/\pi) \\ & \times \left\{[\hat{\sigma}^2_{S_1^w}(t_j) + \hat{\sigma}^2_{S_2}(t_j) - 2\hat{\sigma}_{S_1^w}(t_j)\hat{\sigma}_{S_2^w}(t_j)][\hat{\sigma}^2_{S_1^w}(t_{j'}) + \hat{\sigma}^2_{S_2^w}(t_{j'}) - 2\hat{\sigma}_{S_1^w}(t_{j'})\hat{\sigma}_{S_2^w}(t_{j'})]\right\}^{\frac{1}{2}}\end{aligned} \quad (4.16)$$

To compare two shared-path treatment strategies, the proposed test statistic is

$$\Delta_s^{*w} = \frac{\Delta^w - \hat{E}_s(\Delta^w)}{\sqrt{\widehat{\mathrm{Var}}_s(\Delta^w)}}. \tag{4.17}$$

When the covariances are zero, then this statistic reduces to (4.13). This statistic is also approximately normally distributed with mean 0 and variance 1. Large values of $|\Delta_s^{*w}|$ lead to the rejection of the null hypothesis.

4.6 Simulation studies

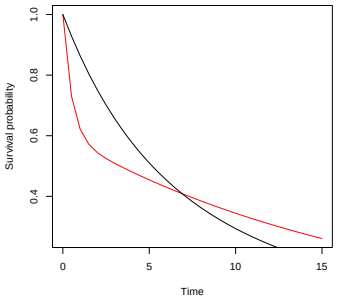
To evaluate the performance of the weighted Lin and Xu test, simulation studies were conducted to study the type I error and the statistical power of the new method. Data was generated from a two-stage SMART design with two first stage treatments and two second stage treatments. The first stage treatment indicator, X_i, was generated from a Bernoulli distribution with $P(X_i = 1) = 0.5$. The response or consent indicator, R_i, was generated from a Bernoulli distribution with $P(R_i = 1) = \pi_r$, where $\pi_r \in (0.4, 0.6)$ to achieve a response rate of 40% and 60%. T_{j0i}, $j = 1, 2$ was generated from an exponential distribution with mean θ_{j0} when $R_i = 0$ and when $R_i = 1$, the time to response, T_{ji}^R, $j = 1, 2$, was generated from an exponential distribution with mean θ_j^R. The second stage treatment indicator, Z_i, was generated from a Bernoulli distribution with $P(Z_i = 1) = 0.5$, equal randomization of responders between B_1 and B_2. The time from response to an event, T_{jki}^*, $j, k = 1, 2$, was simulated from an exponential distribution with mean θ_{jk}, $j, k = 1, 2$. For responders the total survival time is $\tilde{T}_{jki} = T_{ji}^R + T_{jki}^*$, $j, k = 1, 2$. Of interest are the time to event variables, T_{jki} where $T_{jki} = (1 - R_i)T_{j0i} + R_i\tilde{T}_{jki}$, $j, k = 1, 2$. The observed survival time is given by

$$\begin{aligned} T_i = X_i &\left[(1 - R_i)T_{10i} + R_i\left\{Z_i\tilde{T}_{11i} + (1 - Z_i)\tilde{T}_{12i}\right\}\right] \\ &+ (1 - X_i)\left[(1 - R_i)T_{20i} + R_i\left\{Z_i\tilde{T}_{21i} + (1 - Z_i)\tilde{T}_{22i}\right\}\right]. \end{aligned}$$

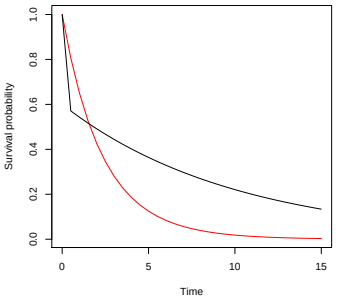
To account for right censoring, C_i was generated from a uniform distribution, $U(0, v)$ and v set to give the desired censoring rate. We defined the observed time as $U_i = min(T_i, C_i)$ with the event indicator $\delta_i = I(T_i \leq C_i)$.

4.6.1 Power estimation

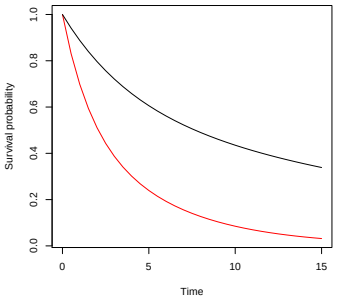
Different scenarios of the alternative distributions were considered. For illustration purposes we considered A_1B_1 and A_2B_1 for independent path treatment policies, for shared path treatment policies we considered A_1B_1 and A_1B_2. Comparison of other treatment sequences can be done similarly. We set α to be 0.05 in all simulations and 1000 iterations were done, we estimated the power as the proportion in the 1000 samples in which the null hypothesis was rejected at 5% level of significance. Below are graphical displays of our simulation scenarios.



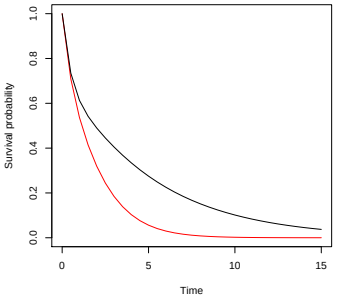
(I) Situation 1



(II) Situation 2



(III) Situation 3



(IV) Situation 4 (shared path)

FIGURE 4.1: Survival curves for A_1B1 (black), A_2B_1 (red) for separate-path treatment policies. A_1B1 (black) and A_1B_2 (red) for shared-path treatment policies. Response rate is 60%

Situation 1: *Crossing survival curves (lower)*

In Situation 1, we have late crossing of the survival curves. The null hypothesis is $H_0 : S_{11}(u) = S_{21}(u)$. To simulate the data for this scenario, the parameters were set as follows: $\theta_{10} = 0.5$, $\theta_{20} = 4$, $\theta_1^R = 3$, $\theta_2^R = 3$, $\theta_{11}^* = 15$, $\theta_{12}^* = 3$, $\theta_{21}^* = 9$, $\theta_{21}^* = 3$. We set $c = (90, 45, 22, 8)$ to achieve 5%, 10%, 20% and 40% censoring for the response rate of 0.6. For a response rate of 0.4, we set $c = (95, 50, 25, 9)$. Figure 4.1 (a) shows the

survival curves for the alternative hypothesis for this scenario. The two survival curves cross at some point at the lower end of the graph. In all our simulations, the sample size is 300.

TABLE 4.1: Simulation results for Situation 1

π_r	Censoring rate(%)	WLX	WLR
0.4	5	1.000	0.231
	10	0.993	0.368
	20	0.989	0.709
	40	0.975	0.998
0.6	5	0.927	0.151
	10	0.906	0.051
	20	0.903	0.092
	40	0.890	0.778

Table 4.1 shows the results for this scenario. In this scenario of late crossing of the survival curves, the WLX test has better power in detecting the difference in the survival curves than the WLR test. For a response rate of 0.4, the power of the WLR test increases with an increase in the censoring rate. Increasing the censoring rate affects the area at the lower end after the two survival curves have crossed. With 40% censoring rate, the power for the WLR test is 0.998 and the power for the WLX test is 0.975, both tests have good statistical power when the censoring rate is around 40%. The power of the WLX test decreases with an increase in the censoring rate. When the response rate is 0.6, the WLX test maintains better power than the WLR test. The statistical power of the WLR test is much reduced when the response rate is 60%. For 10% censoring, the power for the WLR test is 0.051 compared to 0.906 for the WLX test.

Situation 2: *Crossing survival curves (upper)*

In this scenario, we have early crossing of survival curves. The data was generated as follows; $\theta_{10} = 3$, $\theta_{20} = 0.05$, $\theta_1^R = 1$, $\theta_2^R = 3$, $\theta_{11}^* = 1$, $\theta_{12}^* = 3$, $\theta_{21}^* = 7$, $\theta_{21}^* = 3$. For right censoring, we set $c = (47, 25, 12, 3)$ to obtain about 5%, 10%, 20% and 40% censoring for the response rate of 0.6. When $\pi_r = 0.4$, we set $c = (60, 33, 15, 5)$. The results of this simulation are shown in Table 4.2.

TABLE 4.2: Simulation results for Situation 2

π_r	Censoring rate(%)	WLX	WLR
0.4	5	0.999	0.072
	10	1.000	0.136
	20	0.998	0.552
	40	0.986	0.751
0.6	5	1.000	0.928
	10	1.000	0.909
	20	1.000	0.875
	40	0.998	0.104

In this scenario, the WLX test has better power than the WLR test. There is a slight impact on the power of the WLX test when the response rate is increased from 0.4 to 0.6. The power of the WLR test increases with an increase in the censoring rate for 40% response rate. The opposite happens when the response rate is 60%. One explanation for this is the nature of the survival curves for the two response rates. For a response rate of 0.6, the survival curves are as shown in the figure above. Not shown here, the areas between the survival curves are a bit bigger for the response rate of 40% and the crossing point is a bit lower. High censoring rate has an effect to both tests when the response rate is 60%. The WLX test has the power of 0.998 whilst the WLR test has a much reduced power of 0.104. The power of the WLR test decreases with increase in censoring rate .

Situation 3: *proportional hazards*

In this scenario, we consider survival curves that do not cross, that is, where the hazards are proportional. The data was generated as follows: $\theta_{10} = 2$, $\theta_{20} = 2$, $\theta_1^R = 1$, $\theta_2^R = 1$, $\theta_{11}^* = 1$, $\theta_{12}^* = 3$, $\theta_{21}^* = 8$, $\theta_{21}^* = 3$. For right censoring we set $c = (60, 33, 16, 6)$ to obtain about 5%, 10%, 20% and 40% censoring for the response rate of 0.6. When $\pi_r = 0.4$, we set $c = (50, 30, 13, 6)$. The results of this simulation are shown in Table 4.3.

Table 4.3: Simulation results for Situation 3

π_r	Censoring rate(%)	WLX	WLR
0.4	5	1.000	0.998
	10	1.000	0.997
	20	0.999	0.989
	40	0.987	0.920
0.6	5	1.000	1.000
	10	1.000	0.995
	20	1.000	1.000
	40	0.999	1.000

In this scenario, the results suggest that there is no much difference in the statistical power of the two methods. Both methods have good statistical power to detect differences in the survival curves. When the response rate is 40%, both methods have decreasing power as the censoring rate increases. For 40% censoring the WLX test has power of 0.987 and the WLR test has power of 0.920. This is a reduction in power compared to when the censoring rate is 5%. Increasing the response rate from 40% to 60% leads to increase in statistical power in the two methods, the increase in power is more evident for the WLR test. The WLX test has better power than the WLR test only in the case of 10% censoring rate (1.000 and 0.995 respectively). In general, there is no difference in the performance of the two methods in terms of the statistical power when the survival curves do not cross.

Situation 4: *shared path*

We consider the comparison of two shared path treatment policies, A_1B_1 and A_1B_2. Shared path treatment policies share the same non-responders and as such they are not independent. To account for this lack of independence, we modify the variance to include the covariance term. To simulate data for this scenario, we set: $\theta_{10} = 0.5$, $\theta_{20} = 0.5$, $\theta_1^R = 1$, $\theta_2^R = 1$, $\theta_{11}^* = 1.5$, $\theta_{12}^* = 5$, $\theta_{21}^* = 1.5$, $\theta_{21}^* = 1.5$. For right censoring we set $c = (23, 10, 5, 2)$ to obtain about 5%, 10%, 20% and 40% censoring for the response rate of 0.6. When $\pi_r = 0.4$, we set $c = (27, 13, 7, 3)$. The results of this simulation are shown in Table 4.4.

TABLE 4.4: Simulation results for Situation 4

π_r	Censoring rate(%)	WLX	WLR
0.4	5	0.991	0.960
	10	0.990	0.917
	20	0.927	0.742
	40	0.498	0.243
0.6	5	0.998	0.961
	10	0.990	0.934
	20	0.972	0.858
	40	0.690	0.453

The results of this scenario suggests that both methods have good statistical power when the censoring rate is low. When the censoring rate is increased, both methods suffer in terms of power. In the case of 0.4 response rate, the power of the WLX test is 0.498 and the power of the WLR test is 0.243 when the censoring rate is 40%. This is also the case when the response rate is 0.6, the power of the WLX test is 0.690 whilst the power of the WLR test is 0.453. The decline in statistical power of the two tests is severe when the response rate is 0.4.

4.6.2 Type I error estimation

To evaluate the type I error, we simulated null distributions that the treatment policies (A_1B_1 and A_2B_1) are equal. 1000 datasets were generated as follows $\theta_{10} = 0.5$, $\theta_{20} = 0.5$, $\theta_1^R = 3$, $\theta_2^R = 3$, $\theta_{11}^* = 6$, $\theta_{12}^* = 6$, $\theta_{21}^* = 6$, $\theta_{21}^* = 6$. For right censoring we set $c = (60, 35, 20, 6)$ to obtain about 5%, 10%, 20% and 40% censoring for the response rate of 0.6. When $\pi_r = 0.4$, we set $c = (65, 25, 10, 3)$. The type I error was measured as the proportion in which the null hypothesis of equality was rejected at 5% level of significance in the 1000 simulated datasets. The results are shown in Table 4.5 below.

TABLE 4.5: Simulation results for type I error

π_r	Censoring rate(%)	WLX	WLR
0.4	5	0.059	0.059
	10	0.066	0.054
	20	0.060	0.055
	40	0.014	0.053
0.6	5	0.049	0.042
	10	0.053	0.041
	20	0.026	0.038
	40	0.010	0.050

The simulation study indicates that the WLR test has better type I error compared to the WLX test. The type I error for the WLR test are closer to the nominal level. The WLX test generally has its type I error inflated when the response rate is 0.4, the type I error of the WLX test is 0.066 whilst the type I error rate of the WLR is 0.054 for the censoring rate of 10%. Increasing the censoring rate affects the type I error of the WLX test. For a response rate of 0.6 and 40% censoring rate, the type I error of the WLX test is 0.014 compared to 0.053 for the WLR test. The WLX test becomes conservative when the censoring rate is high, this is also true for the response rate of 0.4.

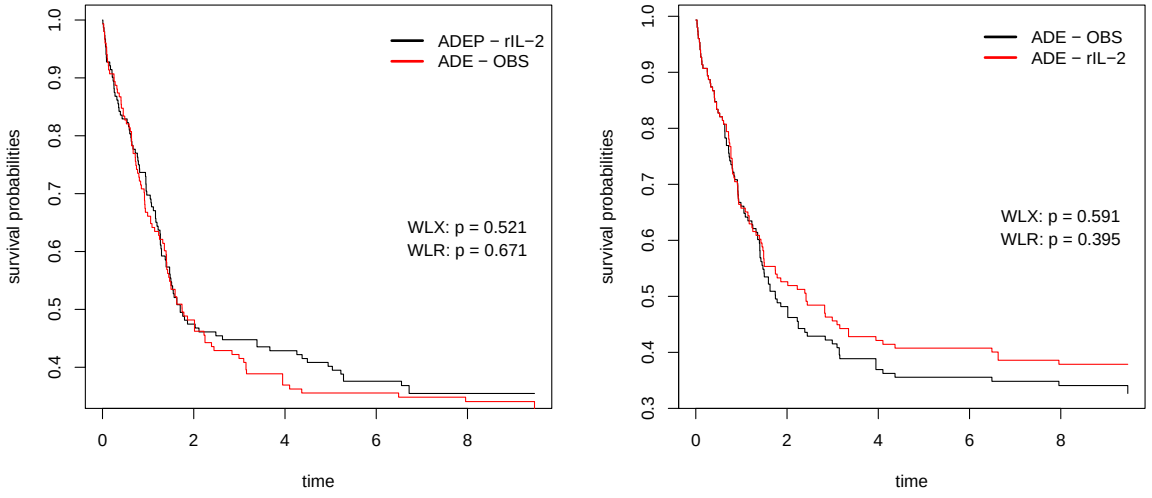
4.7 Application

We applied the WLX test to the CALGB 19808 study. There are four treatment regimes in this study. The survival curves for these four policies are estimated using the WRSE estimator. Table 4.6 summarizes the results of applying the WLX and WLR tests to the clinical trial data. For each pairwise comparison, we report the test statistic and p-value obtained using the two methods. For the shared path analysis, we extended the Lin and Xu test to accommodate dependent samples. The two testing methods give rise to the same conclusions concerning the treatment policies. No treatment policy leads to a better overall survival. Eventhough the analysis in the two papers regarding this trial was done separately for each stage, similar conclusions could be reached even when the analysis focused on the entire treatment sequences as is done here.

TABLE 4.6: Application to CALGB 19808 Study

	WLX		**WLR**	
H_0	Test statistic	P-value	Test statistic	P-value
ADEP-OBS =ADEP-rIL-2	0.605	0.545	1.181	0.237
ADEP-OBS = ADE-OBS	1.262	0.206	0.383	0.702
ADEP-OBS =ADE-rIL-2	0.718	0.473	0.974	0.330
ADEP-rIL-2 = ADE-OBS	0.642	0.521	0.425	0.671
ADEP-rIL-2 = ADE-rIL-2	0.863	0.388	0.157	0.875
ADE-OBS = ADE-rIL-2	0.537	0.591	0.851	0.395

We show in a graph the treatment policies ADEP-OBS versus ADE-OBS and ADE-OBS versus ADE-rIL-2. It can be seen that in the first graph, the survival curves cross many times. The WLX test gives $p = 0.521$ and for the WLR test $p = 0.671$. In cases of crossing survival curves, the WLX test has better power in detecting differences. The second figure is for shared-path treatment policies. Both tests give a non significant result which is appropriate when looking at the graphs.



(I) ADEP-IL-2 vs ADE-OBS (II) ADE-OBS vs ADE-IL-2

FIGURE 4.2: Survival curves for the treatment policies, p refers to p-value.

4.8 Conclusion

In this chapter, we proposed the use of WLX test in comparing two treatment policies. The treatment policies may either be shared paths or independent paths treatment policies. The survival curves are estimated using the WRSE of Guo and Tsiatis (2005). To be able to compare shared paths treatment policies, we modified the Lin and Xu test by incorporating the covariances between the two treatment strategies. In cases where the survival curves cross, the WLX test has better power than the WLR test. Where the survival curves do not cross, the WLR test performs better than the WLX test though the differences are minimal. The type I error for the WLX is a bit inflated. Simulations studies in a study by Li et al. (2015) also revealed that the type I error for Lin and Xu test is inflated. More research is needed in this regard.

Chapter 5

Analyzing safety data for two-stage randomization designs

The safety of patients is an important aspect in the development of new pharmaceutical products. Any biologically active pharmaceutical product is meant to produce benefit to its users but can potentially cause harm as well. Of importance in the development of pharmaceutical products is the understanding of how the potential harms can manifest themselves and at what stage these potential harmful effects can be identified. Some pharmaceutical products fail at the development stage because of unanticipated safety issues. Some products pass through the development stage only to be called from the market place because of some undesired side effects that place the patients at serious health risks (Gould, 2015).

Although safety data are the most common and one of the most important types of data collected in clinical trials, in general more emphasis is given to the efficacy data. More methods are developed to analyze efficacy data but less attention is given to safety data, for example, more methodological developments have happened in the analysis of efficacy data for two-stage randomization designs (Guo and Tsiatis, 2005; Lunceford et al., 2002; Lokhnygina and Helterbrand, 2007; Wahed, 2010; Kidwell and Wahed, 2013) and to our knowledge no study has focused on the analysis of safety data from these designs. There is need to develop sound statistical methodology that provides an accurate and reliable assessment of drug safety.

An adverse event is any untoward medical occurrence in a patient during the course of a clinical trial. An adverse event can be any unfavorable and unintended sign, symptom or disease temporally associated with the use of a medical product, whether it is related to the medical product or not. Adverse events can be classified into different

categories, and in this study we shall focus on serious adverse events. A serious adverse event is defined as any untoward medical occurrence that; (1) may result in death, (2) is life threatening , (3) requires inpatient hospitalization or prolongation of existing hospitalization, (4) results in persistent or significant disability and (5) is a congenital anomaly (Chow and Liu, 2008).

5.1 Review of some methods

Below we give a brief review of methods used in the analysis of safety data for clinical trials with only one stage of randomization. For trials with two treatment arms, that is, the control and new treatment, we have $g = 1, 2$, where g denotes the treatment groups. In some studies there are more than two treatment groups, so $g = 1, 2...$

5.1.1 Incidence proportions

Analysis of safety data is often done using incidence proportions (IPs). These incidence proportions are only valid summaries under the assumption of similar exposure times in both treatment groups. In most cases this assumption is violated because in some trials the exposure times differ. The crude incidence proportion is defined as the number of patients experiencing the adverse event of interest divided by the total number of subjects in each study group. The IP is calculated as

$$IP_g = \frac{a_g}{n_g},$$

where a_g is the number of patients in treatment group g experiencing at least one serious AE and n_g is the total number of patients in treatment group g. The IPs of two groups can be compared using the risk ratio, that is,

$$\text{Risk Ratio} = \frac{IP_1}{IP_2}.$$

Another way of summarizing adverse events data is by using the incidence rate. The incidence rate (IR) is defined as

$$IR_g = \frac{a_g}{(\text{population-time at risk})_g},$$

where a_g is the number of patients in treatment group g experiencing at least one serious AE and $(\text{population-time at risk})_g$ is the population time at risk of the first serious AE

in treatment group g. The denominator in the above equation is the sum of all patients and the times at risk for the first serious AE. A patient who does not experience an AE contributes his/her follow-up time. The incidence rate ratio (IRR) is calculated as

$$IRR = \frac{IR_1}{IR_2},$$

with IR_g being the incidence rate in group g to experience a serious AE.

5.1.2 Exposure adjusted incidence rate

To accommodate patient exposure times, the exposure adjusted incidence rates (EAIR) is defined as the number of subjects experiencing a serious AE divided by the total exposure time among the patients in the treatment group g;

$$EAIR_g = \frac{a_g}{\sum t_{ig}},$$

where a_g is the number of patients in treatment group g experiencing at least one serious AE and t_{ig} is the subject exposure time for individual i until the occurrence of first serious AE in treatment group g. For a subject with no AE, t_{ig} corresponds to the last follow-up time. This type of incidence rate is a valid statistic for treatment comparison when the incidence rate of a specific event is relatively constant over the study period. We interpret the EAIR as the number of serious AEs occurring in a population per unit time. The difference between the IR and the EAIR is that the denominator in the IR is the sum of all patients and the times at risk for the first serious AE. In the EAIR we sum the exposure times only.

5.2 Adverse events and competing risks

Adverse events data are subject to competing risks.

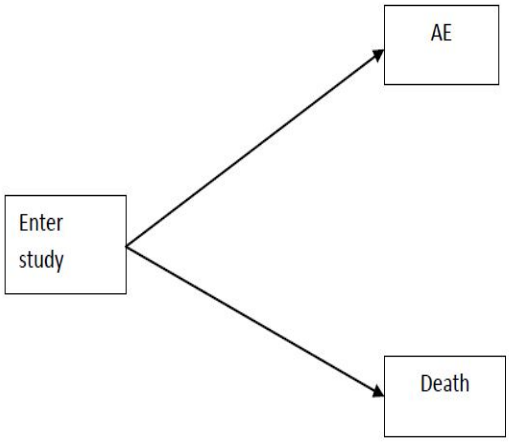


FIGURE 5.1: Competing risks situation for adverse events data.

A patient that enters the study can either experience the AE of interest, die before experiencing the AE or be censored. Since patients may actually die before experiencing the AE, then death is a competing risk for the AE. Figure 5.1 shows the competing events situation. After the patient has died, the AE cannot occur any more. With infinite follow-up and without censoring,

$$\frac{a_g}{n_g} + \frac{d_g}{n_g} = 1,$$

where d_g is the number of deaths without an AE and n_g is the total number of patients in treatment group g.

Consider the time interval $[0, t]$. Without censoring, the probability to experience the composite event (AE or death) is

$$P_g(AE \in [0, t]) + P_g(\text{Death} \in [0, t]) = 1 - P_g(T > t).$$

For estimation, without censoring;

$$\begin{aligned}\hat{P}_g(AE \in [0,t]) + \hat{P}_g(\text{Death} \in [0,t]) &= \frac{a_g^t}{n_g} + \frac{d_g^t}{n_g} \\ &= \frac{a_g^t + d_g^t}{n_g} \\ &= 1 - \hat{P}_g(T > t),\end{aligned} \tag{5.1}$$

where T is the time to the first serious AE or death without an AE, a_g^t is the number of patients in treatment group g experiencing at least one serious AE before or at time t and d_g^t is the number of deaths in group g before or at time t.

5.2.1 Kaplan-Meier estimator

The Kaplan-Meier (KM) estimator is sometimes used to estimate the cumulative incidence function (CIF), $P(AE \in [0,t])$, where death without experiencing an AE is treated as censored observation. In this case,

$$\hat{P}_g(AE \in [0,t]) = 1 - \prod_{u \leq t} \left(1 - \frac{a_g(u)}{r_g(u)}\right), \tag{5.2}$$

where $a_g(u)$ is the number of AEs in treatment group g at u and $r_g(u)$ denotes the number of patients with no AE before u, the so-called risk set, the product is over all AE times.

There has been a number of criticisms in using this approach in estimating $P_g(AE \in [0,t])$. Clearly, it ignores the competing risk set-up that exists in safety data. Another argument against this approach is that $1 - KM_g$ estimator aims at approximating a distribution function which approaches 1 as t becomes larger. On the other hand, $P_g(AE \in [0,t]) + P_g(\text{Death} \in [0,t])$ tends to 1 as t becomes larger, hence the KM based estimator of $P_g(AE \in [0,t])$ is biased upwards (Allignol et al., 2016). Contrary to these arguments, the Kaplan-Meier is still being used in some studies. In defense of this approach, in a response to Schmoor et al. (2016), the authors of the paper (Thanarajasingam et al., 2016) argued that even though the Kaplan-Meier estimator tends to overestimate the $P_g(AE \in [0,t])$, the bias is minimal.

5.2.2 Aalen-Johansen estimator

To estimate the $P_g(AE \in [0,t])$, the Aalen-Johansen estimator should be used in the competing risks situation (Allignol et al., 2016). The cumulative incidence function

(CIF) of an AE is the expected proportion of patients experiencing an AE over the course of time. We note that

$$1 - \hat{P}_g(T > t) = \sum_u \hat{P}_g(T > u^-).\frac{a_g(u) + d_g(u)}{r_g(u)}, \tag{5.3}$$

where $\hat{P}_g(T > u^-)$ is the KM estimator of the probability of not experiencing the composite event AE or death in treatment group g just before time u and $d_g(u)$ is the number of deaths in treatment group g at time u. The sum in (5.3) is the empirical probability to have an AE or death event in $[0, t]$, that is, we are summing over all events times. Now, to get the probability of an adverse event in $[0, t]$, we sum over the empirical probability of experiencing an AE, that is,

$$\hat{P}_g(T \leq t, AE) = \sum_u \hat{P}_g(T > u^-)\frac{a_g(u)}{r_g(u)}, \tag{5.4}$$

we sum over only event times for AEs. Without censoring, (5.4) equals

$$\frac{a_g \in [0, t]}{n_g},$$

this confirms that the incidence proportion is the correct estimate in the absence of censoring.

5.2.3 Hazard functions

In the competing risks situation, a model for the cause-specific hazard function for an AE can be considered. First, we write the total hazard function;

$$\widehat{\alpha}_g(t)dt = \frac{a_g(t) + d_g(t)}{r_g(t)},$$

where g denotes the treatment group. This can be decomposed into the sum of two cause-specific hazards, $\alpha_g^{AE}(t)dt + \alpha_g^D(t)dt$ (D denotes death), which can be estimated by

$$\frac{a_g(t)}{r_g(t)} + \frac{d_g(t)}{r_g(t)}.$$

Having decomposed the hazards in this manner, the Nelson-Aalen estimator of the cumulative hazard to experience an AE is given by

$$\int_0^t \widehat{\alpha}_g(u)^{AE} du = \sum_u \frac{a_g(u)}{r_g(u)}, \tag{5.5}$$

where the sum is over all AE times before t. Only AEs are counted in the numerator of (5.5). In practice, death times are considered as right-censored times. Similarly, for estimating the cumulative hazard function for death, only death events are counted and AE events are censored.

5.3 Analysis of AE data for dynamic treatment regimes

Consider a two-stage randomization design where we have two first stage treatments (A_1, A_2) and two second stage treatments (B_1, B_2). Our interest is in analyzing the AE data for the treatment policies embedded in the SMART design. Different treatment sequences can lead to varying occurrences of adverse events. The aim here is to compare treatment policies in terms of their toxicities or adverse events. To our knowledge, no study has compared treatment policies in terms of their toxicities yet such information can be valuable to many stakeholders involved in the development of personalized medicine. In dynamic treatment regimes, interest is in identifying a treatment strategy that leads to better survival but also such a strategy should be less toxic to the patients.

Allignol et al. (2016) advocate the use of survival analysis methods for analyzing safety data when the primary endpoint in a clinical trial is a time-to-event. In the sequel, we also advocate the use of survival analysis techniques suitable for two-stage randomization designs in the analysis of safety data from these designs. To this end, we propose a methodology to be used in the analysis of safety data from two-stage randomization designs.

Consider a hypothetical experiment where there is no second stage randomization, that is, all patients are assigned to A_1B_1. In such a case, the methods from the previous section apply in the analysis of the safety data. In a two-stage randomization design, we know that some of the patients who could have received B_1 receive B_2 because of the second stage randomization. In the analysis of safety data for two-stage randomization designs, we suggest the use of inverse probability weights since subjects who end up in B_2 are considered missing under A_1B_1. We show how weighting can also be applied in the analysis of safety data for treatment policies. In the literature, two types of weights

have been proposed (Guo and Tsiatis, 2005; Lunceford et al., 2002), in this discussion we shall only focus on time independent weights.

Let $g = 1, 2, ...$ denote the treatment policies. We make the following simplifying assumptions. We note that the events of interest can occur in both stages of the trial and we assume that the AEs occur after response for those who achieve complete remission. This makes the application of the inverse weights to be straight forward. Also, we assume that the states in the competing risk situation are absorbing. Let $W_{i1} = 1 - R_i + R_i Z_i/\pi_z$ be the weight function for A_1B_1, that is, $g = 1$. For A_1B_2, let $W_{i2} = 1 - R_i + R_i Z_i/(1 - \pi_z)$. Similar weights are defined for the treatment policies A_2B_1 and A_2B_2.

5.3.1 Weighted incidence proportions

For treatment policies, we define the incidence proportion as the weighted number of patients experiencing the adverse event divided by the weighted number of subjects in each study group. The weighting is done in such a way that the contribution of a non-responder is given a weight of 1 and a responder is given a weight of $1/\pi_z$ or $1/(1 - \pi_z)$ where π_z is the probability of being randomized to second stage treatment. With this definition,

$$\begin{aligned} WIP_g &= \frac{a_g^w}{n_g^w}, \\ &= \frac{\sum_{i=1}^n W_{ig} I_{ig}(\text{event} = AE)}{\sum_{i=1}^n W_{ig}} \end{aligned} \tag{5.6}$$

where a_g^w is the weighted number of patients in treatment policy g experiencing at least one serious AE, n_g^w is the weighted number of patients in treatment policy g and $I_{ig}(\text{event} = AE) = 1$ if patient i in treatment group g experiences at least one serious AE, it is zero otherwise. As an hypothetical example, we consider a trial where 100 patients are assigned to A_1 and of these 100 patients, 80 respond to the A_1 treatment and are equally randomized between B_1 and B_2. So about 40 patients are randomized to B_1. Suppose that among the responders 15 develop serious AEs and among the non-responders 5 develop AEs. In calculating the WIP, the 5 patients receive a weight of 1 and the 15 patients receive a weight of 2, therefore we have $5 + 30 = 35$. So, $WIP_{A_1B_1} = 35/100 = 0.35$. Without weighting: $IP_{A_1B_1} = 20/60 = 0.33$. In the theory of analyzing dynamic treatment regimes, patients who would have been randomized to B_1 but end up in B_2 are considered missing under the treatment policy A_1B_1. To deal with this 'missingness', inverse weights are used such that we still have 100 patients in the denominator in the above example.

To compare two treatment policies one can use the weighted risk ratio,

$$WRR = \frac{WIP_1}{WIP_2}. \tag{5.7}$$

5.3.2 Weighted exposure adjusted incidence rate

We define the weighted exposure adjusted incidence rate (WEAIR) as the weighted number of subjects experiencing at least one serious AE divided by the weighted exposure time among the subjects in a treatment policy, that is,

$$\begin{aligned} EAIR_g &= \frac{a_g^w}{\sum t_{ig}^w} \\ &= \frac{\sum_{i=1}^{n_g} W_{ig} I_{ig}(\text{event} = AE)}{\sum_{i=1}^{n} W_{ig} t_{ig}}, \end{aligned} \tag{5.8}$$

where a_g^w is the weighted number of patients in treatment policy g experiencing at least one serious AE and t_{ig} is the subject exposure time until the occurrence of first serious AE in treatment policy g. For a subject with no AE, t_{ig} corresponds to the last follow-up time, and W_{ig} is the inverse weight given to individual i in the treatment policy g. To compare two treatment policies one can use the weighted exposure adjusted incidence risk ratio,

$$WEAIRR = \frac{EAIR_1}{EAIR_2}. \tag{5.9}$$

5.3.3 Weighed Kaplan-Meier estimator

Instead of using the usual Kaplan-Meier estimator, we suggest the use of the weighted Kaplan-Meier estimator in analyzing safety data from dynamic treatment regimes. To estimate the probability of an AE in some time interval $[0, t]$, we can use $1 - WKM$, that is,

$$\hat{P}_g(AE \in [0, t]) = 1 - \prod_{u \leq t} \left(1 - \frac{a_g^w(u)}{r_g^w(u)}\right), \tag{5.10}$$

where w denotes that the event and the at risk processes are weighted. The numerator counts the AE events and the denominator gives the number at risk at time u. We weight these processes using the inverse probability weights depending on whether the individual is a responder or non-responder. Deaths before an AE are treated as censored observations. This estimator ignores the competing risks situation that exist in safety data. The most appropriate estimator is based on the Aalen-Johansen estimator.

5.3.4 Weighted Aalen-Johansen estimator

The weighted Kaplan-Meier estimator ignores the competing risks situation that exists in AEs data. Death before an AE is a competing event. For the analysis of AEs data for dynamic treatment regimes, we propose the use of the weighted Aalen-Johansen estimator. The weighted Aalen-Johansen estimator of weighted cumulative incidence function is an appropriate method for estimating the probability of an AE in a competing risks situation:

$$1 - \hat{P}_g(T > t) = \sum_u \hat{S}(u^-)_g^w \frac{a_g^w(u) + a_d^w(u)}{r_g^w(u)},$$

where $\hat{S}(u^-)_i^w$ is the weighted Kaplan-Meier estimator of the probability of not experiencing the composite event AE or death just before time u. We sum over all events times (death or AE). Again, we weight the event processes with inverse weights. The probability of an AE in the time interval $[0, t]$ is given

$$\hat{P}_g(T \leq t, AE) = \sum_u \hat{S}(u^-)_g^w \frac{a_g^w(u)}{a_g^w(u)}, \tag{5.11}$$

where here the sum is over all times of AE before t.

5.3.5 Analysis based on weighted hazards

The all events (AE and death) weighted hazards is given by

$$\widehat{\alpha}_g^w(t)dt = \frac{a_g^w(t) + d_g^w(t)}{r_g^w(t)}.$$

This decomposes into the so-called cause specific weighted hazards, $\alpha_{gAE}^w(t)dt + \alpha_{gD}^w(t)dt$, which can be estimated by

$$\frac{a_g^w(t)}{r_g^w(t)} + \frac{d_g^w(t)}{r_g^w(t)},$$

where $a_g^w(t)$ and $a_g^w(t)$ are the weighted event processes. The quantity $r_g^w(t)$ is the weighted at risk process for treatment policy g.

From the decomposition above, the Nelson-Aalen estimator for the weighted cumulative hazard to experience an AE is

$$\int_0^t \widehat{\alpha}_{gAE}(t)^w du = \sum_u \frac{a_g^w(u)}{r_g^w(u)}. \tag{5.12}$$

In the numerator of (5.12) we only count AEs, that is, we are summing over AEs times. We weight using the inverse probability of being in treatment policy g. In practice, we censor death events before an AE to estimate the weighted cumulative hazard for an AE. The procedure is similar for the weighted cumulative hazard for death without an AE.

5.4 Illustration: CALGB 19808 Toxicity dataset

We illustrate the proposed methodology on the toxicity dataset from the CALGB 19808 study. The efficacy dataset has been described in previous chapters. Several variables were recorded in the toxicity dataset. The adverse events were graded in terms of their severity. The adverse event were graded as mild, moderate, severe, life-threatening and fatal. The adverse event names and their categories are also given. In this illustration, we focus on the analysis of serious adverse events which are called life-threatening (serious AEs) in this dataset. Most of the analysis (other than the incidence proportions and ratios) will be based on the time to the first serious adverse event.

In the development of this methodology, we made some simplifying assumptions. One of them is that we assumed that for responders, the AE occurs after response to the first treatment. This makes it straightforward to apply the inverse probability weights. In this dataset, this assumption is not violated. The efficacy dataset has a variable named *ind_crdays* which gives the number of days from registration to when complete remission was reported. It can be seen that, for almost all the patients who responded, complete remission was achieved very early, for some as early as 24 days. We can then apply the methodology of this chapter assuming the AE occurred after response to the responders to the induction treatments.

In the toxicity dataset, there is not an explicit time to the first serious adverse event. The time is given as an interval made up of two variables which are: AE starting day and AE ending day. The AE ending day refers to the number of days from registration to the end of AE reporting period. The AE occurred in the interval given by the two times. For purposes of this application, we used the AE ending day as our time variable . We could have used the middle value of the interval as our time variable. Interest is in

comparing occurrences of AEs in different treatment policies. To achieve this, we merge the efficacy dataset and the toxicity dataset. The merging was done using the patient number which is present in both datasets.

As described earlier, there are four treatment policies embedded in the CALGB 19808 study, namely; ADE - OBS, ADE - rIL-2, ADEP - OBS and ADEP - rIL-2. In the second stage, some patients were randomized to observation. No active treatment was given to this group as patients were simply observed. There are no adverse events associated with the observation treatment option. In doing the analysis for the AEs, we only considered two treatment policies, which are ADE - rIL-2 and ADEP - rIL-2 for reasons given above. Other than the creation of the time to first serious AE variable, the data was analyzed without any further modifications.

Ignoring the censoring, we calculated the weighted incidence proportions for the treatment policies. The weighted incidence proportion for ADE - rIL-2 is 0.9797 and the weighted incidence proportion for the ADEP - rIL-2 is 0.9615. The probability of having a serious adverse event was slightly higher in the ADE - rIL-2 treatment policy. The weighted risk ratio, $WRR = 0.9797/0.9615 = 1.015$. The estimated risk of experiencing at least one serious AE is approximately the same in the two treatment policies.

To calculate the weighted exposure-adjusted incidence rate, we consider three scenarios a patient might be in during the trial. A patient who experiences a serious AE while still in the exposure time contributes to the time at risk his/her weighted time to the AE. A patient who dies without experiencing an AE contributes to the time at risk his/her weighted time to death. Lastly, a patient who does not experience a serious AE contributes to the time at risk for an AE his/her weighted time to the end of exposure. The weighted time at risk of exposure in the ADE - rIL-2 treatment policy is 9498.593 days and for the ADEP - rIL-2 is 10079.83 days. The WEAIR in the ADE - rIL-2 treatment policy is 0.0128 and WEAIR in the ADEP - rIL-2 is 0.0124. There is no major difference in the WEAIRs for the two treatment policies. This can be shown by calculating the weighted exposure-adjusted incidence risk ratio, $WEAIRR = WEAIR_1/WEAIR_2 = 0.0128/0.0124 = 1.036$. There is no difference in number of serious AEs occurring daily in the two treatment policies.

Figure 5.2 is obtained by treating death as censored and then taking $1-WKM$. The probability of an AE is estimated by $1 - WKM$ and this approach has been criticized as it ignores the competing risks situation. The graph shows no differences in the probabilities of experiencing an AE in the two treatment regimes.

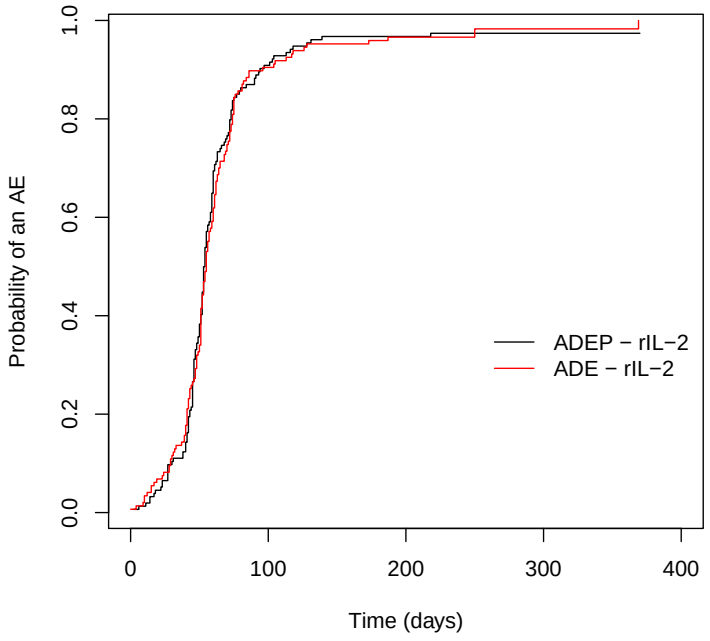


FIGURE 5.2: Estimating the probability of an AE using weighted Kaplan-Meier estimator.

As mentioned before, the most appropriate approach of estimating the probability of an AE is the use of the weighted Aalen-Johansen estimator of the CIF.

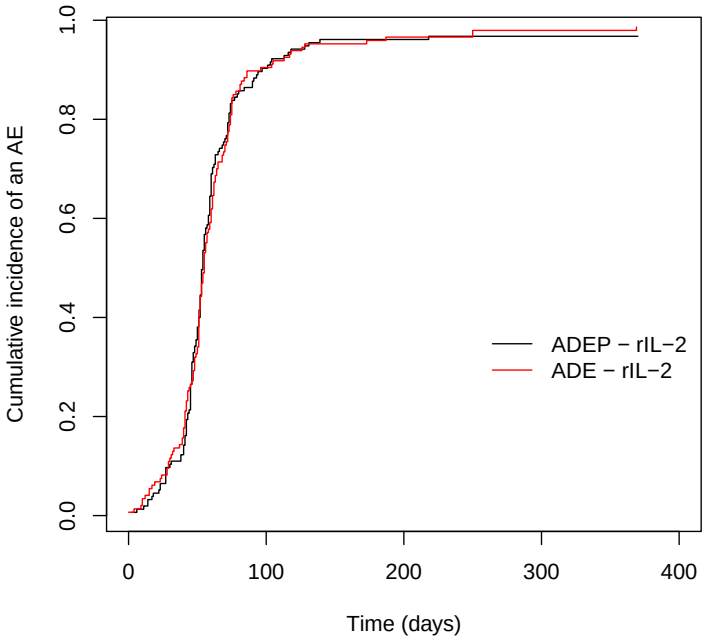


FIGURE 5.3: Estimating the probability of an AE using weighted Aalen-Johansen estimator.

The graphs obtained from the weighted Kaplan-Meier looks similar to the ones from the weighted Aalen-Johansen estimator in Figure 5.3. This is not surprising since there were few deaths in the dataset. One should expect the two estimators to be similar if there are few deaths (competing risks). For this reason we do not show the graph for CIF for death events.

The estimation of the probability of an AE by the weighted Kaplan-Meier tends to overestimate the probability. This could not be shown clearly in this analysis as there were few competing events (deaths). It can be seen though, that the graph of the weighted Kaplan-Meier is slightly above the graph from the Weighted Aalen-Johansen estimator in the tail of the distribution. This is depicted in Figure 5.4.

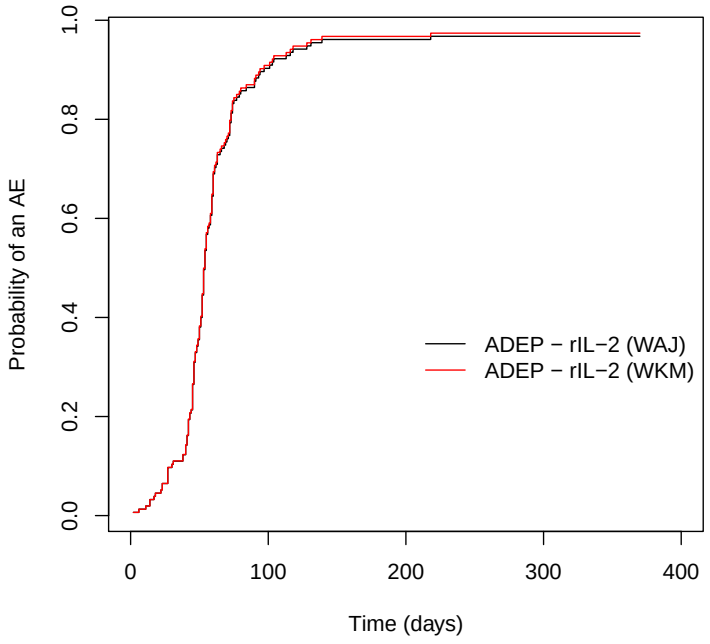


FIGURE 5.4: A comparison of weighted Aalen-Johansen and weighted Kaplan-Meier estimators.

In the reply to Schmoor et al. (2016) by Thanarajasingam et al. (2016) they argued that, even though the Kaplan-Meier estimator is biased upwards, its bias is minimal. This may not be true in all cases, but only in the case where the competing events are few. Since interest is in the probability of a competing event, that is, AE or death, if the count for deaths is close to zero, then the two estimators will be similar.

In doing the analysis based on event-specific hazards, we only report the weighted cumulative hazards for an AE. Due to lower numbers of death before a serious AE, we did not include the graphs for death before an AE. When ignoring the competing event of death, we observe some differences between the cumulative hazards of ADEP - rIL-2 and ADEP - rIL-2. The cumulative hazard of experiencing an AE was higher in the ADEP - rIL-2 than in the ADE - rIL-2 treatment policy in the time period 100 days to about 180 days, thereafter the hazard of an AE is higher in the ADEP - rIL-2 treatment policy. For the earlier times, there is no much difference in the hazards of an AE between the two treatment policies. The cumulative hazard of experiencing an AE was equal in the first 100 days. This is shown in Figure 5.5 below.

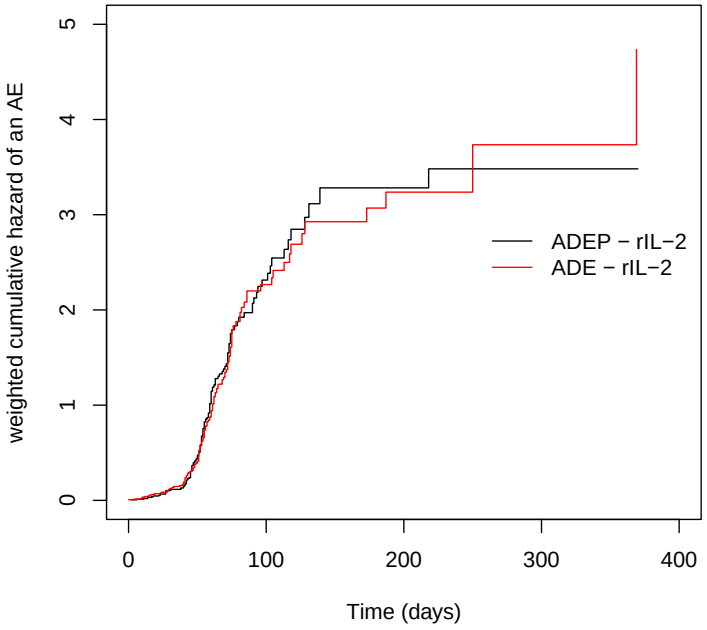


FIGURE 5.5: Weighted cumulative hazards for AEs using the Nelson-Aalen estimator.

5.5 Conclusion

There has been an acknowledgment that safety data does not receive the attention as efficacy data (Gould, 2015). In most cases, the analysis of safety data has been done using crude incidence rates and this type of analysis may not be adequate. The use of time-to-event statistical methods is common practice for efficacy endpoints in clinical studies but such methods are rarely applied in the analysis of safety data. In this chapter, we give a general overview of the methods that are applicable to single stage study with a time-to-event endpoint. We then propose a methodology for analyzing safety data from two-stage randomization designs which uses inverse probability weights. The weighting is done in a similar way as in the analysis of efficacy data. We used time-independent inverse weights. A responder represents $1/\pi_z$ patients who could have potentially been assigned to the treatment policy of interest. A non-responder only represents himself. In doing so, we have made the analysis of the safety data be in sync with the analysis of efficacy data from these study designs. We have focused on the time to the first serious AE.

The important aspect to note in safety data is the presence of competing risks situation. A patient who enters the study can experience the AE of interest, die before

experiencing the AE or be censored. The use of the Kaplan-Meier estimator is not encouraged but if the competing events are few, then the bias is minimal. In the general practice, it is recommended to use the weighted Aalen-Johansen estimator proposed in this chapter.

Chapter 6

Conclusions

6.1 Discussion

Treatment of complex diseases such as cancer, HIV, leukemia and depression usually follows complex treatment sequences. Patients are randomized to first-stage treatments, and upon response, a second randomization to the second-stage treatments is done. In this thesis, we focused on two-stage randomization designs with time-to-event endpoints. For these designs, interest is on estimating survival distributions and comparing different treatment policies.

We started by reviewing statistical methods for estimating survival distributions in two-stage randomization designs. A simulation study was conducted to compare the performance of the three methods. The LDT estimator was found to be affected by high censoring rates and low response rates. All three methods give similar survival probabilities if the censoring rate is low and the response rate is not very low.

Wahed (2010) proposed a parametric approach for estimating survival distributions for standard SMART designs. We extended Wahed's approach to accommodate time-varying SMART designs. Our approach uses the notion of convolution of two random variables. Crossing survival curves can pose a challenge in comparing survival distributions. To remedy this problem in two-stage randomization designs, we proposed a weighted version of the Lin and Xu test (Lin and Xu, 2010). Our simulation studies suggest that the weighted Lin and Xu test has better statistical power to detect differences in treatment policies even if the survival curves cross. In cases where the survival curves do not cross, there is no gain in statistical power in using our approach.

The analysis of adverse events data has not been given the same attention as the analysis of efficacy data. The key point in adverse event data is the presence of the competing risks situation. We developed a methodology for analyzing adverse events

data. We believe that adverse events data should be given the same focus as efficacy data.

6.2 Future directions of research

The parametric approach developed in the second chapter depends on the convolution of random variables. There is a challenge in finding probability density functions for some survival distributions other than the exponential distribution. Future research is needed in this aspect where numerical methods could be used. An R package that can implement the parametric approach could help in making the methodology useful. The weighted Lin and Xu test depends on a fixed value of the correlation coefficient. This restriction can be removed by finding a way of estimating the correlation coefficient. This is one direction we intend to follow in future to remove the dependence on a fixed given correlation coefficient. Also, the weighted Lin and Xu test is at the present restricted to pairwise comparisons, we intend to extend it in the future to be able to compare more than two groups.

There are many directions for future research in the area of analyzing safety data from two-stage randomization designs. In the analysis of adverse events, we assumed that the adverse event of interest occurs after response for the responders. This assumption can be relaxed. One way of doing this is to treat as a non-responder a responder who experiences an adverse event in the first stage.

Bibliography

Allignol, A., Beyersmann, J., and Schmoor, C. (2016). Statistical issues in the analysis of adverse events in time-to-event data. *Pharmaceutical statistics*, 15(4):297–305.

Bessate, A. and El Bouanani, F. (2016). A very tight approximate results of MRC receivers over independent weibull fading channels. *Physical Communication*, 21:30–40.

Brigham, E. O. and Morrow, R. (1967). The fast fourier transform. *IEEE spectrum*, 4(12):63–70.

Cheson, B. D., Cassileth, P. A., Head, D. R., Schiffer, C. A., Bennett, J. M., Bloomfield, C. D., Brunning, R., Gale, R. P., Grever, M. R., and Keating, M. J. (1990). Report of the national cancer institute-sponsored workshop on definitions of diagnosis and response in acute myeloid leukemia. *Journal of Clinical Oncology*, 8(5):813–819.

Chow, S.-C. and Liu, J.-P. (2008). *Design and analysis of clinical trials: concepts and methodologies*, volume 507. John Wiley & Sons.

Collett, D. (2015). *Modelling survival data in medical research*. CRC press.

Dai, T. and Shete, S. (2016). Time-varying SMART design and data analysis methods for evaluating adaptive intervention effects. *BMC Medical Research Methodology*, 16(1):112.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Gould, A. L. (2015). *Statistical methods for evaluating safety in medical product development*. John Wiley & Sons.

Guo, X. (2005). *Statistical analysis in two stage randomization designs in clinical trials*. PhD thesis, Department of Statistics, North Carolina State University.

Guo, X. and Tsiatis, A. (2005). A weighted risk set estimator for survival distributions in two-stage randomization designs with censored survival data. *Int J Biostat*, 1:1–15.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Kidwell, K. M. and Hyde, L. W. (2016). Adaptive interventions and SMART designs: Application to child behavior research in a community setting. *American Journal of Evaluation*, 37(3):344–363.

Kidwell, K. M. and Wahed, A. S. (2013). Weighted log-rank statistic to compare shared-path adaptive treatment strategies. *Biostatistics*, pages 1–14.

Klein, J. P., Logan, B., Harhoff, M., and Andersen, P. K. (2007). Analyzing survival curves at a fixed point in time. *Statistics in medicine*, 26(24):4505–4519.

Kolitz, J. E., George, S. L., Marcucci, G., Vij, R., Powell, B. L., Allen, S. L., DeAngelo, D. J., Shea, T. C., Stock, W., Baer, M. R., et al. (2010). P-glycoprotein inhibition using valspodar (PSC-833) does not improve outcomes for patients younger than age 60 years with newly diagnosed acute myeloid leukemia: Cancer and Leukemia Group B study 19808. *Blood*, 116(9):1413–1421.

Li, H., Han, D., Hou, Y., Chen, H., and Chen, Z. (2015). Statistical inference methods for two crossing survival curves: a comparison of methods. *PLoS One*, 10(1):e0116774.

Lin, X. and Xu, Q. (2010). A new method for the comparison of survival distributions. *Pharmaceutical statistics*, 9(1):67–76.

Lokhnygina, Y. and Helterbrand, J. D. (2007). Cox regression methods for two-stage randomization designs. *Biometrics*, 63(2):422–428.

Lunceford, J. K., Davidian, M., and Tsiatis, A. A. (2002). Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, pages 48–57.

McLachlan, G. and McGiffin, D. (1993). On the role of finite mixture models in survival analysis. *Statistical methods in medical research*, 3(3):211–226.

Miyahara, S. and Wahed, A. S. (2010). Weighted Kaplan–Meier estimators for two-stage treatment regimes. *Statistics in medicine*, 29(25):2581–2591.

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W. E., Gnagy, B., Fabiano, G. A., Waxmonsky, J. G., Yu, J., and Murphy, S. A. (2012). Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychological methods*, 17(4):457.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of educational Psychology*, 66(5):688.

Ruckdeschel, P. and Kohl, M. (2010). General purpose convolution algorithm in s4-classes by means of fft. *arXiv preprint arXiv:1006.0764*.

Schmoor, C., Bender, R., Beyersmann, J., Kieser, M., and Schumacher, M. (2016). Adverse event development in clinical oncology trials. *The Lancet Oncology*, 17(7):e263–e264.

Tang, X. and Melguizo, M. (2005). DTR: An R package for estimation and comparison of survival outcomes of dynamic treatment regimes. *Journal of Statistical Software*, 65(7):1–28.

Thall, P. F., Sung, H.-G., and Estey, E. H. (2002). Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *Journal of the American Statistical Association*, 97(457).

Thanarajasingam, G., Atherton, P. J., Novotny, P. J., Loprinzi, C. L., Sloan, J. A., and Grothey, A. (2016). Longitudinal adverse event assessment in oncology clinical trials: the toxicity over time (ToxT) analysis of Alliance trials NCCTG N9741 and 979254. *The Lancet Oncology*, 17(5):663–670.

Wahed, A. S. (2010). Inference for two-stage adaptive treatment strategies using mixture distributions. *J R Stat Soc: Ser C (Appl Stat)*, 59:1–18.

Sifiso Vilakati

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4174
e-mail: vilakati@stat.unipd.it

Current Position

Since October 2014; (expected completion: October 2017)
PhD Student in Statistical Sciences, University of Padova, Italy.
Thesis title: Inference Following Two-Stage Randomization Designs with Survival Endpoints
Supervisor: Prof. Giuliana Cortese

Research interests

- Two-stage randomization designs
- Two-stage designs

Education

January 2012 – April 2013
Master degree in Statistics.
University of Kwa-Zulu Natal, Faculty of Science, South Africa
Title of dissertation: "Multilevel Modeling of HIV in Swaziland Using Bayesian and Frequentist Approaches. "

Supervisor: Prof. Thomas Achia
Final mark: 2(i)

January 2011 – December 2011
Bachelor degree (Honours) in Statistics.
University of Kwazulu Natal, Faculty of Science, South Africa
Title of dissertation: "Contribution of Media Exposure to Smoking in Swaziland. "
Supervisor: Prof. Thomas Achia
Final mark: Cum laude.

August 2004 – August 2006
Bachelor degree in Statistics and Mathematics.
University of Swaziland, Faculty of Social Science
Title of dissertation: "Statistical Analysis of Choice of Science in High School. "
Supervisor: Mr. Sidumo Masango
Final mark: 2(ii).

Further education

---

23 June 2017
European Statistical Meeting on Analysis of Safety Data in Clinical Trials
Organizing Institution: European Federation of Statisticians in the Pharmaceutical Industry (EFSPI)
Organizer: Astellas, Leiden, Nerthelands
Instructor: several

29 June – 3 July 2015
Course on Multivariate Event Times Data
University of Padova, Italy
Instructor: Thomas Scheike (University of Copenhagen, Denmark)

12 - 15 October 2015
Course on Categorical Data Analysis
University of Padova, Italy

Instructor: Alan Agresti (University of Florida, USA)

Work experience

January – July 2014
University of Swaziland.
Lecturer for Statistics .

February – November 2012
University of Kwa-Zulu Natal.
Tutor for 1^{st} years in Statistics.

February – November 2011
University of Kwa-Zulu Natal.
Tutor for 1^{st} years in Statistics and Mathematics.

July 2007 – September 2010
Ministry of Education, Swaziland.
High School Teacher for Mathematics.

Awards and Scholarship

2014
PhD Scholarship, University of Padova.

2012
Masters Scholarship, University of Kwa-Zulu Natal.

2012
Cum laude Bursary, University of Kwa-Zulu Natal.

Computer skills

- R
- SPSS

- Stata

Language skills

SiSwati: native; English: fluent, Italian: basic

Conference presentations

Vilakati, S., Cortese, G., (2016). Oral presentation. Title: Estimating Survival Distributions for Two-Stage Randomization Designs. *ISCB Annual Conference*, Birmingham, UK, 21 - 25 August 2016.

References

Prof. G. Cortese
University of Padova
via C. Battisti, 241,
35121 Padova
Phone: +39 049 827 4159
e-mail: gcortese@stat.unipd.it

Prof. S. Motsa
University of Swaziland
Private Bag 4,
Kwaluseni, M201
Phone: (+268) 2518 4011
e-mail: sandilemotsa@gmail.com